Answers to comments

<table>
<tr><th colspan="3">Reviewer 1</th></tr>
<tr><td colspan="3">Major comments</td></tr>
<tr><td>1</td><td>List parameters all the parameters used for all bioinformatics tools, like Blastp and CD-Hit, MAFFT, and hmmsearch.</td><td>Section 4.1 (General methods for building CAZy families) has been removed and the details have been moved to the sections for each family. Specific parameters have been added where they were missing.</td></tr>
<tr><td>2</td><td>Data Availability. I could not find the submitted data on CAZY. Maybe I am looking in the wrong place or searching for the wrong terms (GTxx1!).</td><td>The CAZy database does not assign family numbers before formal acceptance to avoid conflict with other families being created at the same time (as we know all too well, the time between submission and acceptance is highly variable). This is why placeholders GTxx1-GTx17 are used. The final family numbers will be provided at proof stage unless we can update our paper after final acceptance and before the production of the proofs (dear Editor, please let us know).</td></tr>
<tr><td>3</td><td>In the methods, add justification for thresholds, where applicable. For eg, L378: threshold of 110? Or L381, threshold of 160</td><td>The following was added:

Line 543-547:
“A series of networks were built using different bit score thresholds. The members of the resulting SSN clusters were identified using NetworkX [50] and MSAs of the members were built with MAFFT v7.508 using the L-INS-i strategy. The MSAs were inspected using Jalview [51], 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).”

Line 550-554:
“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</td></tr>
</table>

		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)”
4	Section 4.3: Is the tree rooted? If so, then what is the outgroup? If it’s not rotted, who are the inference about clades being made?	Phylogenetic trees generated using Aclust are ‘mid-point rooted’. Using an outgroup would add one additional root node to the tree, but would not change the tree topology within the branch containing the initial sequence set. Clade subdivision at a given node can be inferred from the tree by comparing the branch length above (in the direction of tree root) and sum of branch lengths below (in the direction of tree leaf nodes). A detailed description has been added in section 4.1 (Line 481-494).
Minor comments		
5	Figure 7 caption: Which structures are alphafold indicate them, e.g. 2.	The sentence in the caption has been changed to: “The secondary structures were taken from the crystal structures for family GTxx1 and GTxx3 (6BAR and 7TPG respectively) and from AlphaFold models for all other families.”
6	Section 4.9 is redundant to the second paragraph of section 4.8	The section has been deleted.
7	L90: What is high sequence conservation? Provide a number/estimate.	The sentence has been changed to “To date this new family contains over 4800 members with sequence identity greater than 38% over 414 residues” (Line 156)
8	L96: Italicize vide infra	Has been changed to “see below”
9	L138-L141, L341-342: Wrong opening quotes	The quotes have been changed. Furthermore, quotes are only used the first time we refer to superclusters.
10	L167-169: What is this bacteria? Is the genome available? If so, can two different polymerases be verified with colocalization of other genes	We checked the original paper from which we got this sequence (Hu, 2013), and they mention that they were uncertain about the

	involved in the pathway?	serotype of this strain. We have deleted this BP-Pol sequence from the dataset. This affects Figure 2, Figure 5 and Supplementary Figure 4. We have added the following sentence in section 4.6: “The BP-Pols for *A. baumannii* O7 and O16 were omitted, because of uncertainty of their serotypes” (Line 531-532) As for the second family with inconsistency in stereochemistry, the sugar that is different comes from *P. aeruginosa* O4 (polymerase accession: AAM27782.1 from Islam, 2014). In this case, we were not able to solve the issue. There is no available genome for the serotype. We have kept this BP-Pol in the dataset.
11	L364: see section “General methods used for building CAZy families” -> see section 4.1, similar to L367	This has been deleted (see answers to comment 1)
12	L375: 95->95% sequence identity; Missing period	Has been changed to: “95% identity.” (Line 538)
Reviewer 2		
Major comments		
13	In order to compare the activities of different BP-Pols, the authors assign predicted activities to sequences based on carbohydrate information in the CSDB database. However, we have some doubts regarding the validity of this approach. The idea of assigning activities to Wzy homologues in a genome based on reported glycan structures seems to be based on a very idealistic assumption. How can one be sure that the Wzy homologue in question corresponds to the particular glycan structure that was reported? Is the database fully accurate in distinguishing e.g. O- and K- antigens? Some O-antigens are built on K-antigen cores. Furthermore, many of these entries could be based on outdated or misinterpreted information. It also seems unclear how one can	CAZy families are usually based on characterized founding members. However, the BP-Pols are very tricky to express, and there is only one that has been experimentally characterized (Woodward, 2010). Therefore, we decided to rely on the BP-Pol sequences that have been reported in seven studies that have annotated BP-Pol sequences in the polysaccharide gene clusters, which we believe are reliable. The corresponding glycan structures have also been reported in these studies, except for *E. coli,* for which the glycan structures are reported in a separate paper. For *Streptococcus pneumoniae* there were eight serotypes for which the sugar unit was unknown at the time of the review paper, but where the structure has been elucidated

	be confident in determining which linkage is formed by the Wzy homologue in question. It would be better to rely only on Wzy homologues for which the activity have actually been characterized.	later. We used CSDB to find these eight structures and checked the references. We thus only used CSDB to find the relevant literature and to retrieve the SNFG figures and linear strings. As for the linkages formed by the polymerase, this has been determined for most of the sugars in the studies mentioned above. This is based on prediction of the initial sugar in the unit. A few changes have been made in the dataset after checking all the references again. Two BP-Pols have been deleted (from A. baumannii O7 and Pseudomonas aeruginosa O15). We have added a detailed description of all this in section 4.5 and 4.6 (Line 526-601).
14	Molecular phylogeny, sequence clustering, and HMMs obviously rely on accurate multiple sequence alignments (MSAs). In this instance, the highly divergent sequences of BP-Pols (and other GT-C sequences) were difficult to align, which is why the authors developed several intelligent sequence-based strategies to overcome this. However, the most definitive sequence alignments can often be produced from structural alignments, as structure is usually more highly conserved over sequence. With the advent of AlphaFold, this may be achievable. The authors could make more use of this tool to help validate their sequence alignments and structural predictions; after all, prediction of structure from classification seems to be one of the key points of the paper. Would it be possible to align a set of representative AlphaFold structures to gain more insight? For example, aligning the more distantly related members of each new family, or comparing family representatives within the same clan? Furthermore, do the SSN/HHblits clusters correspond to any clusters within the Foldseek database? Of course, this comes with the limitation that the AlphaFold predictions may be biased in some way.	Structure-based clustering is a very powerful method, but it is computationally heavy for large datasets. We find that our sequence-based approach gives good results and is much less computationally demanding. Regarding Foldseek, we checked the 17 representative sequences from Figure 7, and found that they are all in different Foldseek clusters except for the GTxx9 and GTx10 representatives which are in the same Foldseek cluster (cluster A0A1D7ZNT6). This indicates that the Foldseek clusters compare with our clusters, although they are not completely equivalent. Regarding structural superimpositions of AlphaFold models, we have included superimpositions of five distantly related members of GTxx4 in Supplementary Fig. 3. We have also included superimpositions of family representatives within the same clans in Supplementary Figs. 5-7.

Minor comments		
15	Line 28 – shouldn’t the higher limit for number of TMs be at least 14, not 13? E.g. ALG6 has 14 TMs	Yes, thank you ! The change has been made and a second reference has been added. (Line 110)
16	Lines 33 and 273: in several places, it is claimed that stereochemistry is conserved within each GT family. However, this is not always the case, e.g. the GT8 family has both inverting and retaining activities. So therefore, the precedence for all family members conserving the same mechanism is not as strong as it is made out (relevant for later arguments).	The stereochemistry of glycosyl transfer is well conserved within the families of GTs previously established in the CAZy database. However, as the reviewer points out, one exception is found in family GT8, where 53 enzymes have been experimentally characterized with 52 (~98%) retaining the anomeric configuration and one inverting it. The latter is a LPS β-1,3-N-acetyl-glucosaminyltransferase of Helicobacter pylori (PMID=15814825). No other inverting GT has been characterized so far in family GT8, suggesting that the inverting GT8 enzymes may be restricted to Helicobacter pylori. Thus outside a very small number of exceptions, the stereochemical outcome of the reaction of CAZy GT families is well conserved. Our present work shows that, like for GTs families with GT-A or GT-B fold, the stereochemical outcome of glycosyl transfer is also well conserved for families with the GT-C fold. We have amended the text to account for possible exceptions (line 49): “With almost no exceptions, this feature is conserved in previously defined sequence-based families”
17	Line 151: The anomeric carbon carrying the Und-PP moiety need not be ‘α’ – for most L-sugars, for example, it would be ‘β’.	This has been changed to: “The repeat unit structures are always axially linked (α for D-sugars and β for L-sugars) to the Und-PP moiety before polymerization.” (Line 246-247)
18	Line 192: ‘analyzes’ should presumably be ‘analyses’.	This has been corrected (Line 305).
19	Line 214: ‘on the outside of the membrane’ – does this mean on the outer face of the membrane, or excluded from the transmembrane region? Or both?	Has been changed to: “the conserved residues are located on the outer face of the membrane.” (Line 329)

20	Line 198-219 and elsewhere, please ensure consistent formatting, for example super- vs. subscript ‘GT-CBx’	This has been corrected.
Reviewer 3		
Major points		
21	My main concern with the paper in its current form it is challenging to understand the main points of the paper, due to the inclusion of only peripherally related text and analysis. The paper would be greatly improved if it were more focused on the main points. I had to read the paper closely and several times to figure out what the main points were, and I think they are approximately the following: a. CAZy did not previously cover a distinct subclass of bacterial glycosyltransferases, called GT-CB, which are all membrane proteins with 10 transmembrane helices. b. This paper attempts to expand the number of GT-CB families in CAZy. c. GT-CBs fall into four large subgroups of functions, peptidoglycan polymerases, enterobacterial common antigen polymerases (ECA-Pol), O-antigen ligases (O-Lig), and other bacterial polysaccharide polymerases (BP-Pol). d. The first three subgroups form well defined families of proteins that are similar enough in sequence be aligned in a multiple sequence alignment, while the BP-Pol group is highly diverse in sequence, and cannot be aligned. e. A strategy was developed to identify families within the BP-Pol group, which involved manual review and adjustment of sequence similarity networks (SSNs) to define small families of highly similar sequences, followed by using HMM comparisons to merge small families together into a single, larger family when they have very similar amino acid sequence profiles. f. Using this strategy, the authors created 14 families of BP-Pols, which cover approximately 40% of the BP-Pols of known substrate specificity. The remainder form small groups which may be	We have done a major rewrite of the abstract and introduction to include these points. We have also tried to make the story clearer throughout the results section. The background on peptidoglycan polymerases has been moved from section 2.1 to the introduction.

	expanded and added to CAZy in the future as more bacterial genomes are sequenced. g. An important advance is that the authors identify four distinct subgroups within GT-CB, each of which conserves the overall reaction product stereochemistry. I would suggest focusing on these (or similar) points in the abstract, so that the main points of the paper are clear to readers. The first two sentences of the current abstract are not directly related and can be removed, as well as mention of a similarity score (see below).	
22	In a few specific places in the manuscript, text and figures that are not directly related to these points should be shortened or removed. Some specific suggestions are: 1. The focus of the introduction should be background on the CAZy database, and background specifically on the GT-CB class, such as the previous work of Alexander and Locher to identify that class (this is now in section 2.6), and the four functional subgroups in point c above. All the text about glycans in general, and glycosyltransferases in general, can be greatly abbreviated and point to publications with further information.	The mention of the work of Alexander and Locher has been moved to the introduction, and the text about glycans in general has been abbreviated.
23	2. The analysis of similarity score of substrate (carbohydrate monomer unit) structure within each protein family was not very informative, and could be removed. This would include text in section 2.5 starting with line 170, as well as Figures 4 and 5, and section 4.7. Specifically, the authors did not describe how the similarity score was developed, and the authors point out that a separate ms in in preparation, which would be appropriate. The blocks in Figure 4 do not correlate extremely highly with family (there are many 0 scores in each block), and the SNFG diagrams in Figure 5 are not overwhelmingly similar graphically. The figures in supplementary figure 3 are much more informative than the	We have carefully considered the point raised by this (and not by the others) reviewer and have decided to maintain the parts on the similarity score of glycan structures as, even imperfect, it does show a trend and, to our knowledge, there is no method currently available to quantify glycan similarity. We feel it is worth reporting this. However, to make this section more convincing, we have considerably amended our description by adding more details on what inspired the method and on its principles. We have removed Figure 5 and instead included one of the trees from

	similarity scores, and the authors can use this as support for the statement that each new family they have defined tends to share elements of substrate specificity in common.	supplementary figure 3 in the main text (now Figure 3).
24	Consistent with these suggestions, the authors might consider a more descriptive title for the paper, something like: Increasing the coverage of sugar-diphospholipid-utilizing glycosyltransferase families in the CAZy database	We considered the reviewer’s proposal, but prefer our original title, which fits better the findings that are about GT diversity more than merely adding families to the CAZy database.
25	Line 100. Aclust. The URL for obtaining this program should be provided section 4 to ensure reproducibility.	The following has been added in section 4.1 (Line 482): “Source code may be obtained via GitHub at https://github.com/GarryGippert/Aclust).”
26	There is no description in section 4 of how the authors constructed the GTxx1 family, and this should be added.	In the results (section 2.1), we have added this: “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 was created and initially populated by using BLAST against GenBank, and subsequently by searching against GenBank with an HMM built from the retrieved sequences.” (Line 129-132) In the methods section we added a new section 4.3 (Building the peptidoglycan polymerase family (GTxx1)) (Line 495-501).
27	Line 434. The authors claim the new families are on the CAZy website, but I could not find them there. It is important that these new families are made available on the website before the manuscript is published. Also, I assume the family identifiers/names in the paper that contain x’s are placeholders, like GTxx1, and if so the names of the families in the manuscript will	The CAZy database does not assign family numbers before formal acceptance to avoid conflict with other families being created at the same time (as we know all too well, the time between submission and acceptance is highly variable). This is why placeholders GTxx1-GTx17 are used. The final family numbers will be provided at proof stage

	be updated to be the same as on the CAZy website prior to publication.	unless we can update our paper after final acceptance and before the production of the proofs (dear Editor, let us know).
Minor comments		
28	Please use the full name for each subgroup (see point c. above) the first time it appears in the text, with the abbreviation in parentheses.	Done.
29	Line 80, ‘sequence-wise we found excellent sequence similarity’, please rephrase. Sequence-wise is an awkward phrase, and sequence similarity should be quantified, e.g. greater than % sequence identity over x residues”.	The sentence has been deleted and we write instead: “with a pairwise sequence identity of 19% over 221 residues for the most distant members” (Line 133)
30	Line 81, ‘57,200’. The authors should specify the source of the sequences—are these all the related sequences in GenBank, or some other source?	We have specified “GenBank members” (Line 132)
31	4. Line 96 and other places, ‘vide infra’ and ‘vide supra’. The standard English ‘see below’ and ‘see above’ would be preferable.	Has been changed to ‘see below’ and ‘see above’.
32	Line 113 ‘area’. Do you mean clade? Or subtree?	We have changed the text to: “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 form one clade in the tree, except for a few outliers.” (Line 180-182)
33	Line 162-169 and Figure 2 legend, axial and equatorial. Most of the ms distinguishes between inverting and retaining, can you use that same terminology here for simplicity? If not, please explain.	We have added the following clarification of the relation between axial/equatorial and retaining/inverting: “The repeat-unit structures are always axially linked (alpha for D-sugars and beta 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. If the bond

		formed by the polymerase is axial, then the mechanism is retaining and if the bond formed by the polymerase is equatorial, then the mechanism is inverting.” (Line 246-250) In the Figure 2 legend we have written: “Equatorial bonds are shown in orange, implying an inverting mechanism. Axial bonds are shown in blue, implying a retaining mechanism.”
34	Figure 7 should somehow indicate which structures are experimental, and which are predicted by alphafold	The sentence in the caption has been changed to: “The secondary structure was retrieved from the crystal structures for family GTxx1 and GTxx3 (6BAR and 7TPG respectively) and from AlphaFold models for all other families.”
35	Lines 290, 302, ‘folding’. Suggest ‘fold’ instead	We have changed it to ‘fold’
36	Line 425 ColabFold, is there a reference that could be added?	The reference has been added (Line 628).

Updates to figures

Figure 1

1. We show singletons in both networks.
2. The clusters are coloured to better show the relation between the two networks.
3. There are small changes in the network, due to the update of the dataset.

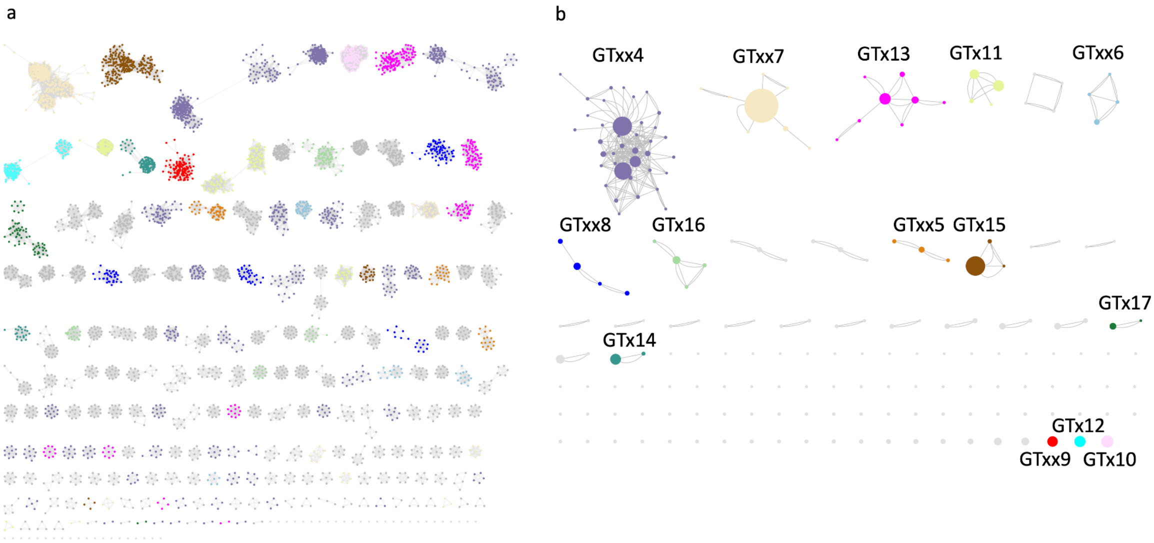


Figure 2

The figure was updated after one seed was removed as explained in comment 10. All the bonds made by the polymerase in GTxx4 are now equatorial.

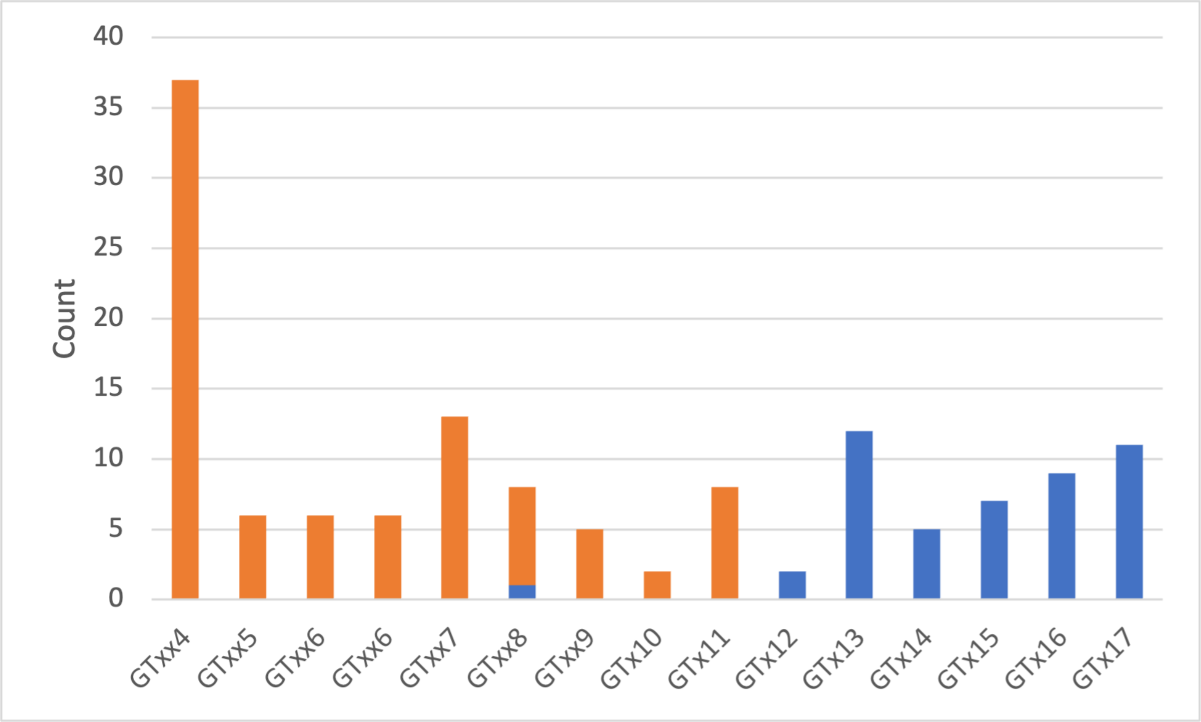


Figure 3

This figure is new. What was figure 5 before has been replaced by this figure showing a phylogenetic tree with BP-Pols and their corresponding sugar structures. See answers to comment 23.

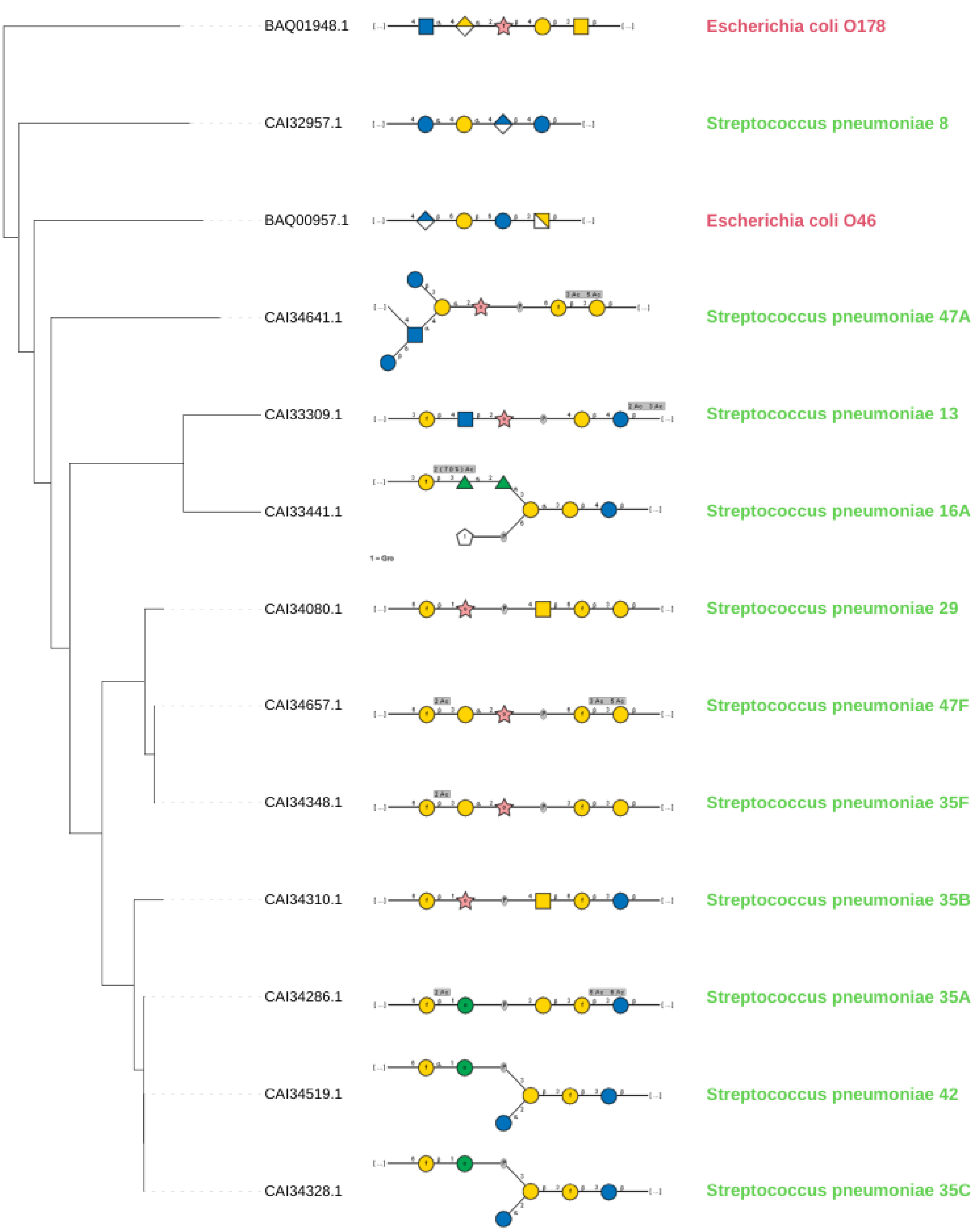

BAQ01948.1
Escherichia coli O178
CAI32957.1
Streptococcus pneumoniae 8
BAQ00957.1
Escherichia coli O46
CAI34641.1
Streptococcus pneumoniae 47A
CAI33309.1
Streptococcus pneumoniae 13
CAI33441.1
Streptococcus pneumoniae 16A
1 = Gro
CAI34080.1
Streptococcus pneumoniae 29
CAI34657.1
Streptococcus pneumoniae 47F
CAI34348.1
Streptococcus pneumoniae 35F
CAI34310.1
Streptococcus pneumoniae 35B
CAI34286.1
Streptococcus pneumoniae 35A
CAI34519.1
Streptococcus pneumoniae 42
CAI34328.1
Streptococcus pneumoniae 35C

Figure 4

What was Figure 3 before is now Figure 4. No changes have been made to this figure.

Figure 5

What was figure 4 before is now Figure 5. There are few changes as a result of the sugar corrections (see answers to comment 13).

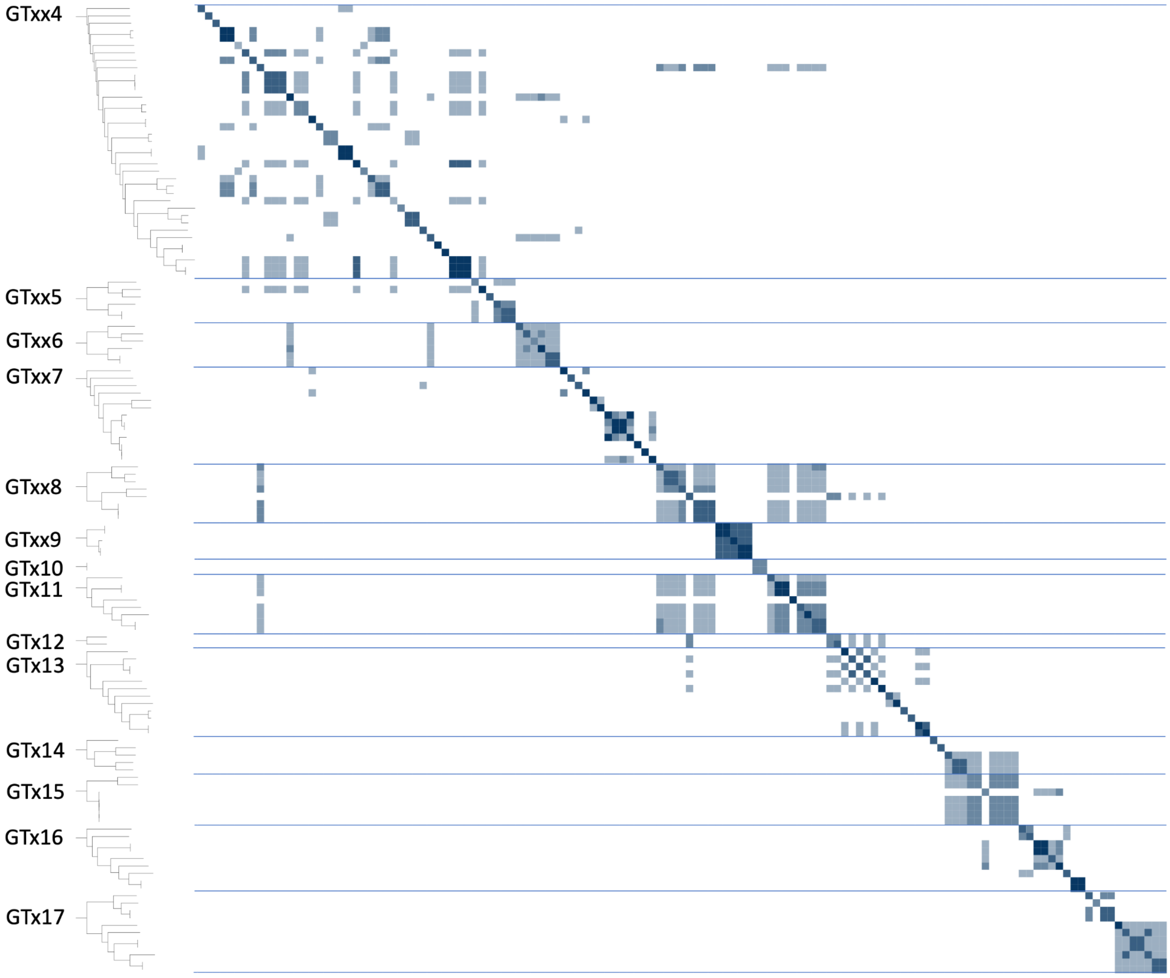

GTxx4
GTxx5
GTxx6
GTxx7
GTxx8
GTxx9
GTx10
GTx11
GTx12
GTx13
GTx14
GTx15
GTx16
GTx17

Figure 6

“BP-Pol” has been added to the labels on the x- and y-axis.

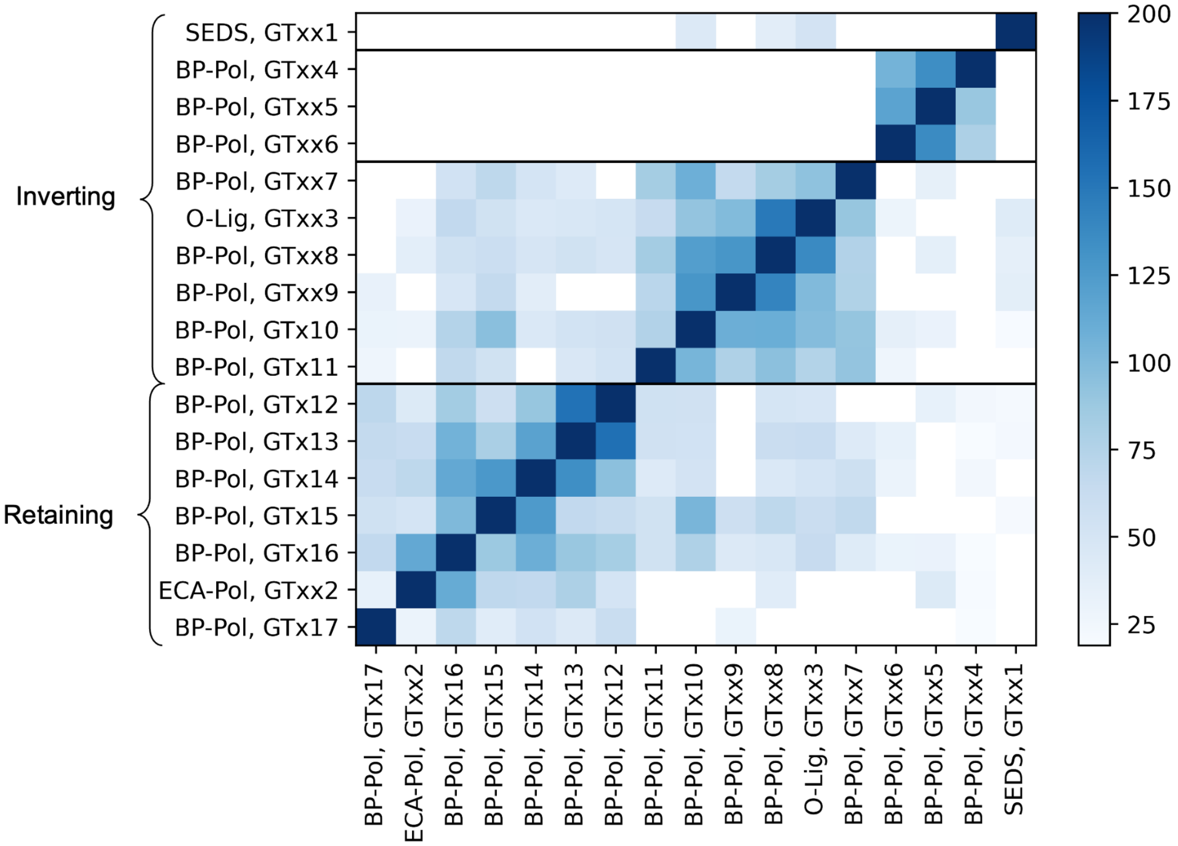


Figure 7

There have been a few corrections in the conserved residues. Residues that are conserved in more than 98% of the sequences are shown.

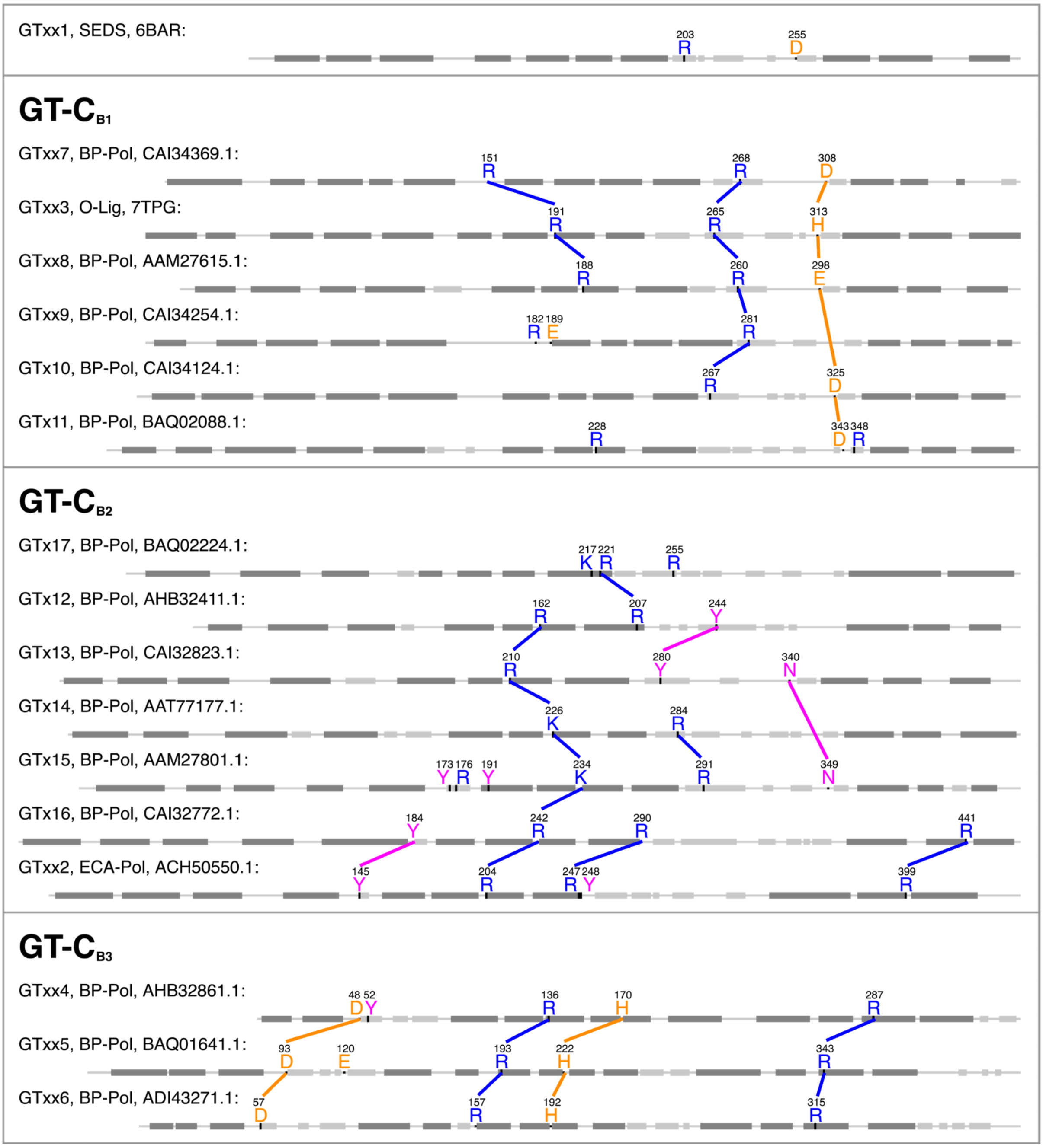
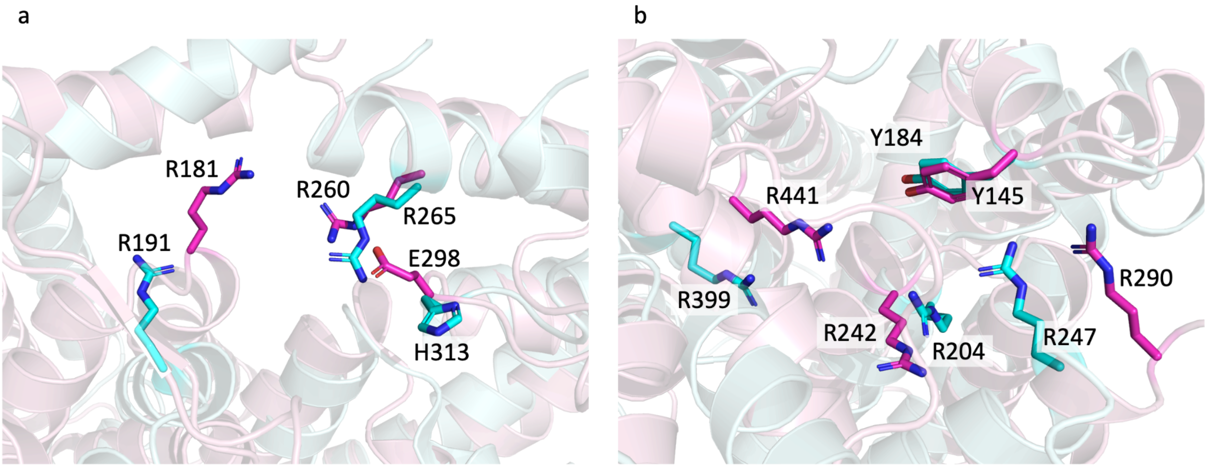

GTxx1, SEDS, 6BAR:
203 R
255 D
GT-C$_{B1}$
GTxx7, BP-Pol, CAI34369.1:
151 R
268 R
308 D
GTxx3, O-Lig, 7TPG:
191 R
265 R
313 H
GTxx8, BP-Pol, AAM27615.1:
188 R
260 R
298 E
GTxx9, BP-Pol, CAI34254.1:
182 189 R E
281 R
GTx10, BP-Pol, CAI34124.1:
267 R
325 D
GTx11, BP-Pol, BAQ02088.1:
228 R
343 348 D R
GT-C$_{B2}$
GTx17, BP-Pol, BAQ02224.1:
217 221 K R
255 R
GTx12, BP-Pol, AHB32411.1:
162 R
207 R
244 Y
GTx13, BP-Pol, CAI32823.1:
210 R
280 Y
340 N
GTx14, BP-Pol, AAT77177.1:
226 K
284 R
GTx15, BP-Pol, AAM27801.1:
173 176 Y R
191 Y
234 K
291 R
349 N
GTx16, BP-Pol, CAI32772.1:
184 Y
242 R
290 R
441 R
GTxx2, ECA-Pol, ACH50550.1:
145 Y
204 R
247 248 R Y
399 R
GT-C$_{B3}$
GTxx4, BP-Pol, AHB32861.1:
48 52 D Y
136 R
170 H
287 R
GTxx5, BP-Pol, BAQ01641.1:
93 D
120 E
193 R
222 H
343 R
GTxx6, BP-Pol, ADI43271.1:
57 D
157 R
192 H
315 R

Figure 8

Residue labels have been added.



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

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

Updates to supplementary figures

Supplementary Figure 1

No changes

Supplementary Figure 2

No changes

Supplementary Figure 3

This figure is new. It shows structural superimpositions of AlphaFold models of distantly related members of BP-Pols from family GTxx4.

Supplementary Figure 4

This figure was Supplementary Figure 3 before. Species and serotype have been added

Supplementary Figures 5-7

These figures are new. They show structural superimpositions of representative members from each family in the three clans.