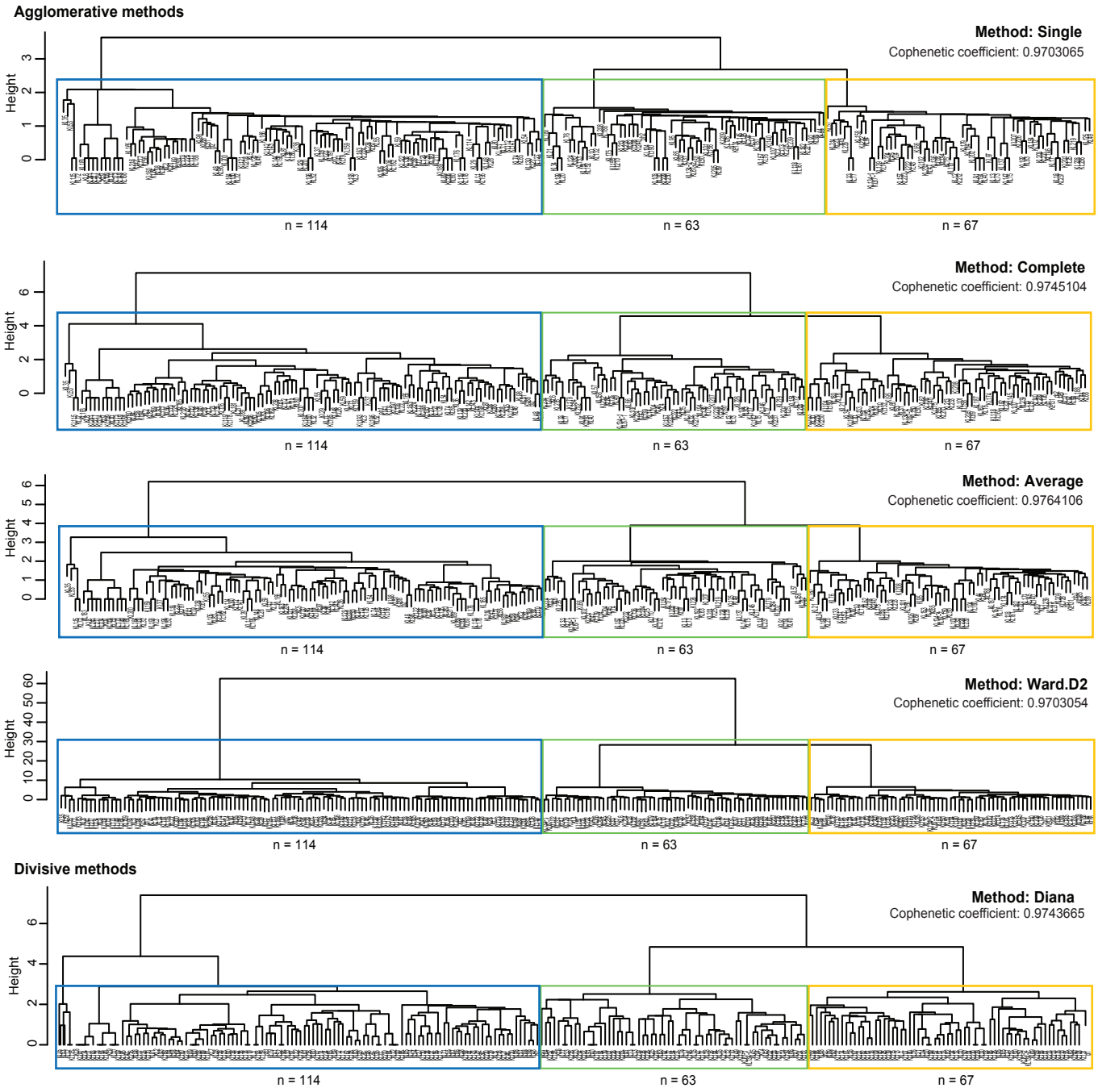
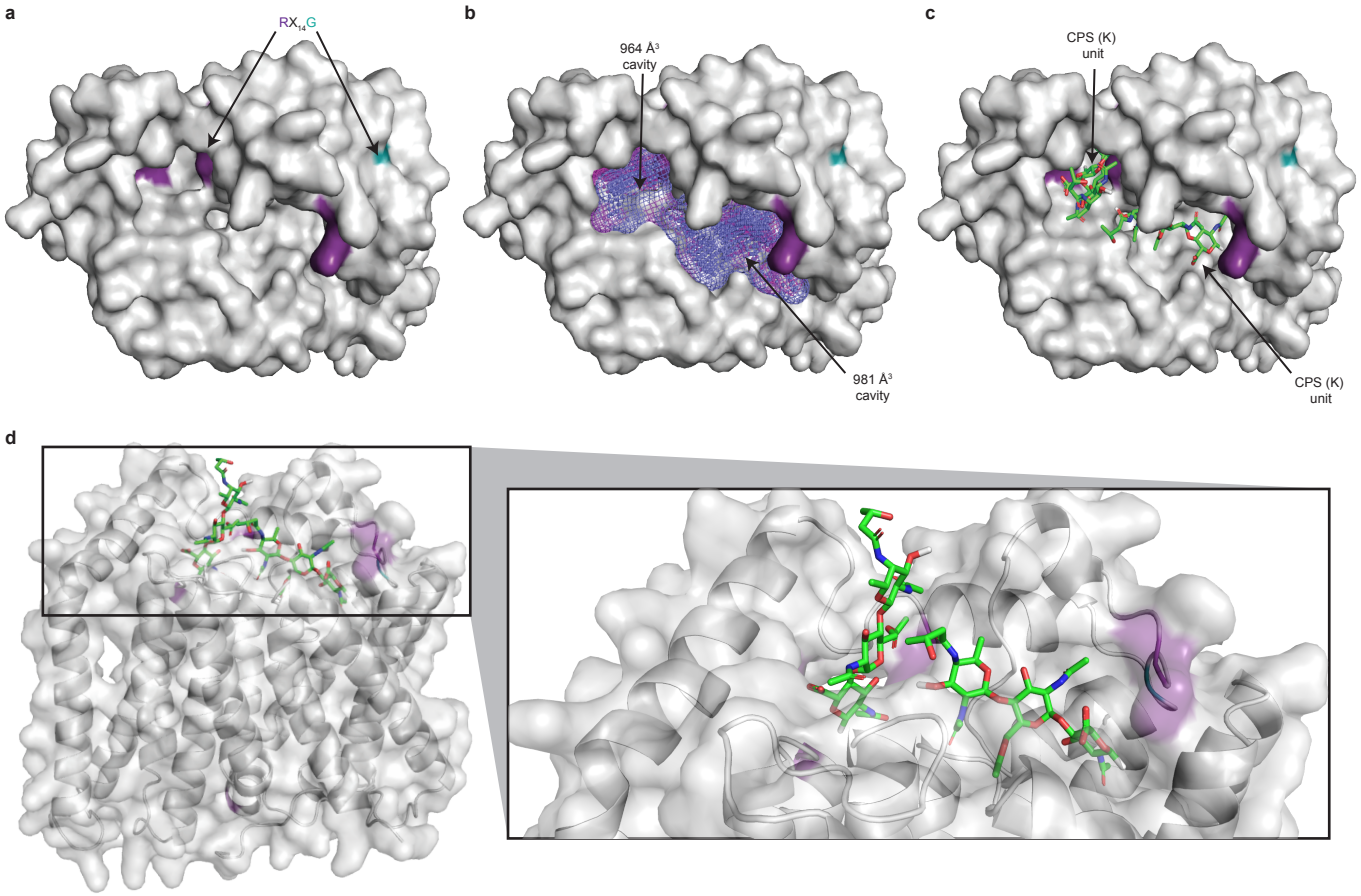
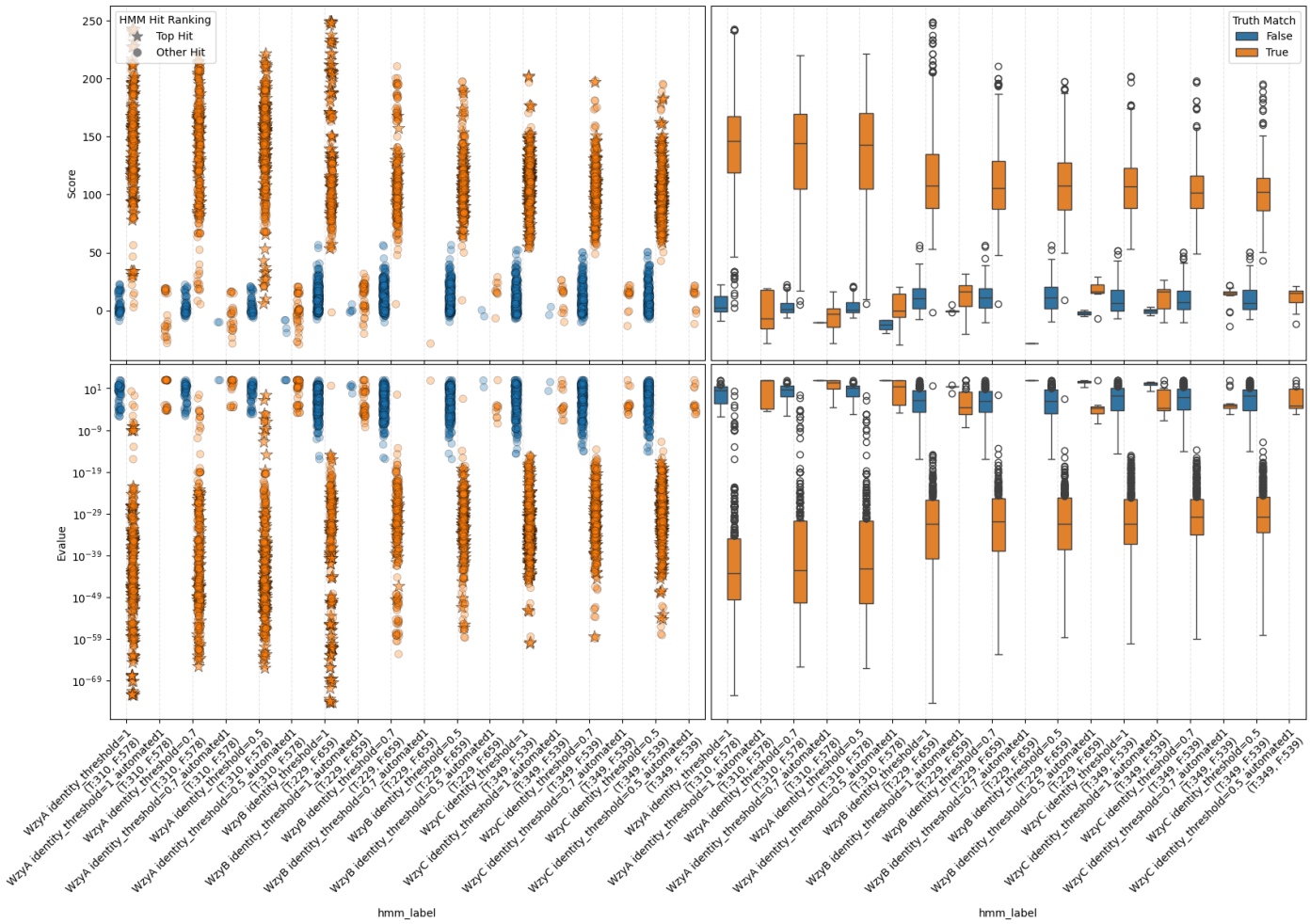


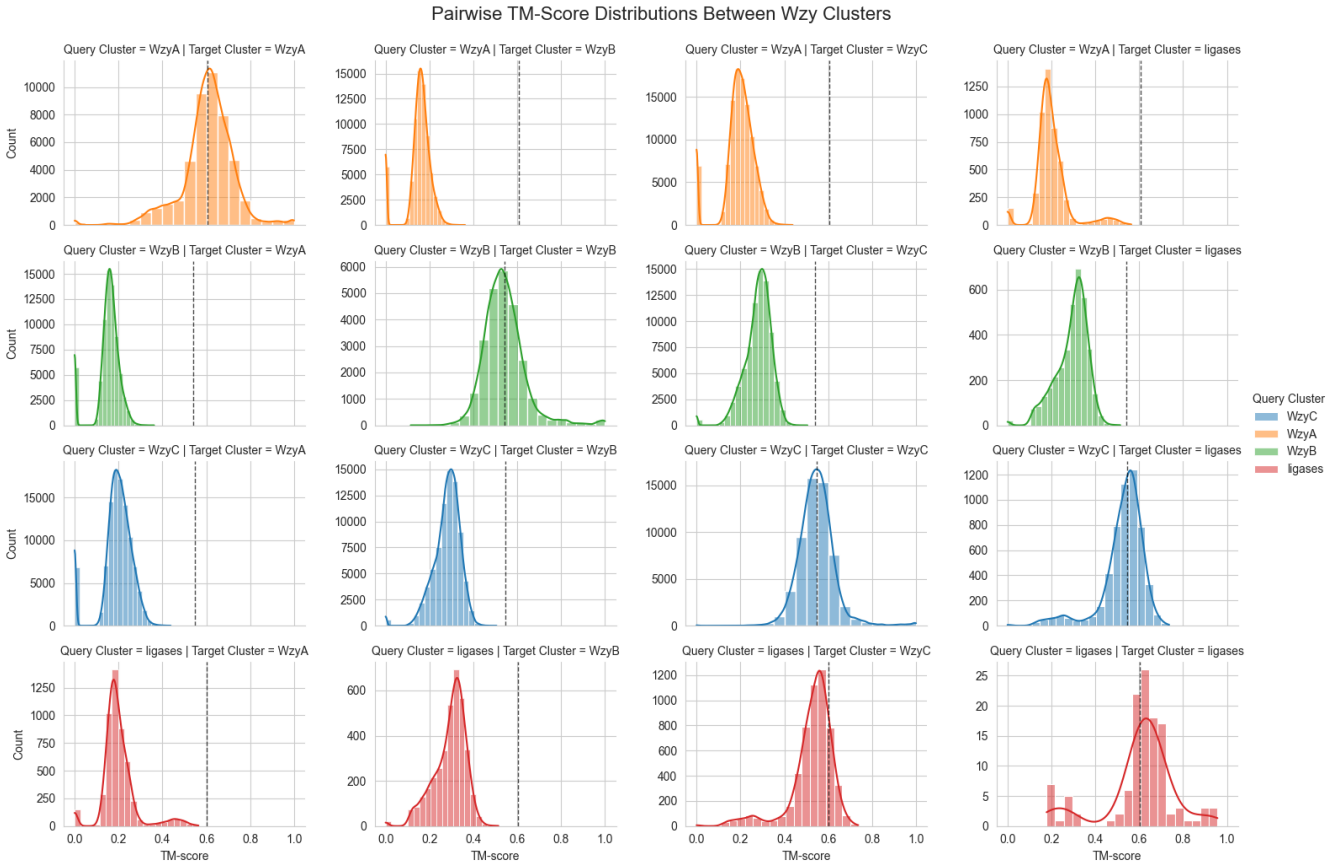
Extended Data Figure 1. Comparison of hierarchal clustering methods using tertiary structures of *A. baumannii* Wzy polymerases. Comparison of dendrogram visualisations and cophenetic coefficient calculations of agglomerative (single, average, complete, ward.D2) and divisive (DIANA) clustering methods. Colour boxes indicate subgroups when dendogram is cut into three.



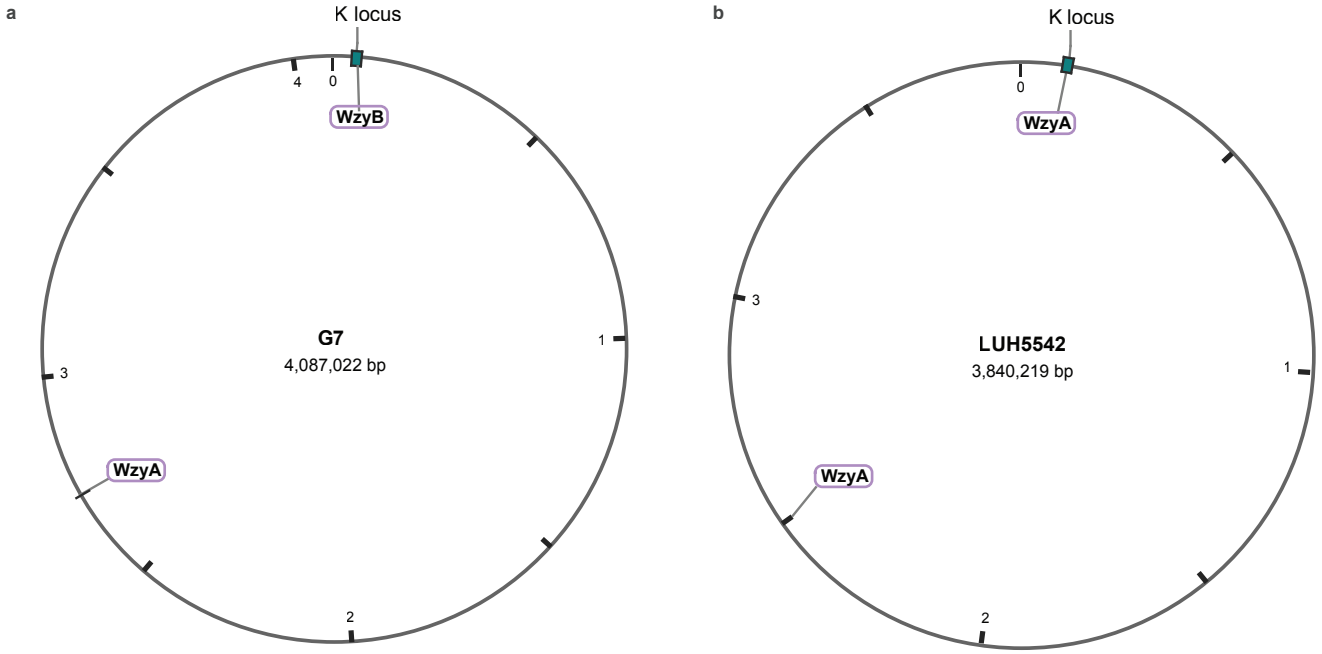
Extended Data Figure 2. Wzy modelled in complex with glycan substrates. (a) Top view indicating relevant arginine (R; purple) and glycine (G; green) residues. $RX_{14}G$ motif in PL2 indicated. **(b)** CavitOmix cavity predictions. **(c)** Structure modelled and complexed with two K1 glycans, which were each placed in a predicted periplasmic cavity. **(d)** Lateral view of Wzy (transparent surface) complexed with K1 glycan, showing zoomed-in region of K1 glycan placement in cavities.



Extended Data Figure 3. Bitscore (top row) and e-value (bottom row) distributions of custom HMMs to WzyA, WzyB and WzyC ground truths. HMM profiles tested are those built from MSAs trimmed row-wise at varying identity thresholds of 0.5, 0.75 or 1 (i.e. no trimming), or those followed by trimming column-wise using the "automated1" heuristic algorithm. Colours indicate whether they match the ground truth (orange vs blue) and shapes indicate if matches were the top hit across all HMMs tested (star vs circle). Distributions are visualised as data points (left) and boxplots (right).



Extended Data Figure 4. Pairwise TM-score distributions for structural alignments between proteins that belong to the same cluster (diagonal) or different clusters. The grey dotted line represents the mean pairwise TM-score for protein alignments along the diagonal.



Extended data Figure 5. Chromosome maps of *A. baumannii* isolates G7 (a) and LUH5542 (b). Sequences drawn to scale from NCBI accession numbers CP175642.1 (G7) and CP195996.1 (LUH5542). Location of the K locus in each chromosome is indicated by a green box. Positions of detected Wzy are shown. Megabase coordinates are shown inside the circles.

Extended Data Table 1. Analysis of the major loops in WzyA, WzyB and WzyC sequences

Family	Loop	Length range (aa)	Median length (aa)	Mean length (aa)	Enriched amino acids	% sequences with			
						R	G	H	RX_nG
WzyA	PL1	23-70	50	50.2	Y,W,F,N,M,I,L,D,A	81.0	94.0	42.6	41.2
	PL2	11-67	30	30.4	Y,W,M,N,A,L,I,S,F	44.3	90.0	24.2	17.8
WzyB	PL1	7-38	22	23.1	M,A,Y,N,E,I,L,D,S,G,F,C	65.5	99.6	28.4	11.8
	PL2	70-138	98	98.6	Y,F,N,I,S,M,L,W,G,D	92.1	100	67.2	32.8
	CL1	7-29	16	16.5	M,N,I,C,K,A,E,L,Y	-	-	-	-
WzyC	PL1	8-55	30	28.8	M,F,Y,N,I,E,L,A,S,D	81.7	93.4	39.0	49.6
	PL2	51-108	75	75.5	M,F,Y,N,I,E,L,A,S,D	94.6	100	67.1	92.3

Extended Data Table 2. Sensitivity and specificity of Wzy HMM profile searches

HMM Set	N Seqs	N Wzy	TP	FN	FP	TN	Sensitivity	Specificity
Pre-existing	8431	394	340	54	1	8036	0.862944	0.999876
Novel	8431	394	393	1	4	8033	0.997462	0.999402

TP = true positive; FN = false negative; FP = false positive; TN = true negative