Capsular specificity in temperate phages of *Klebsiella pneumoniae* is driven by diverse receptor-binding enzymes

Aleksandra Otwinowska[1,*], Janusz Koszucki[2,3,*], Vyshakh R. Panicker[2,3], Jade Leconte[2], Sebastian Olejniczak[1], Kathryn E. Holt[4,5], Edward J. Feil[6], Eduardo P. C. Rocha[7], Bogna Smug[2], Barbara Maciejewska[1], Zuzanna Drulis-Kawa[1†] & Rafal J. Mostowy[2,†]

1) Department of Pathogen Biology and Immunology, University of Wroclaw, Wrocław, Poland

2) Malopolska Centre of Biotechnology, Jagiellonian University, Kraków, Poland

3) Doctoral School of Exact and Natural Sciences, Jagiellonian University, Kraków, Poland

4) Department of Infection Biology, London School of Hygiene & Tropical Medicine, London, United Kingdom

5) Department of Infectious Diseases, School of Translational Medicine, Monash University, Melbourne, Victoria, Australia

6) The Milner Centre for Evolution, Department of Life Sciences, University of Bath, Bath, UK

7) Institut Pasteur, Université Paris Cité, CNRS, UMR3525, Microbial Evolutionary Genomics, Paris, France

∗ These authors contributed equally to this work.

† Correspondence: rafal.mostowy@uj.edu.pl, zuzanna.drulis-kawa@uwr.edu.pl

Abstract

In bacteriophages infecting *Klebsiella pneumoniae*, capsule specificity is a major determinant of host range due to the presence of capsule-specific depolymerases. Yet for temperate phages, the genetic and functional basis of this specificity remains less well understood. Depolymerases appear unexpectedly rare in prophage genomes, raising unresolved questions about which prophage genes mediate capsule tropism, whether this apparent scarcity reflects biological or ecological differences versus annotation limits, and whether prophage-encoded receptor-binding proteins (RBPs) are functionally active. To address these questions, we analysed 3,900 *Klebsiella* genomes from diverse ecological niches to identify prophage-encoded proteins mediating capsule tropism. We conducted a genome-wide association study (GWAS) correlating prophage protein clusters (from 8,105 prophages) with confidently assigned bacterial K-loci. GWAS identified high-confidence predictors for 16 out of 35 most diverse K-loci, of which 14 were receptor-binding proteins (RBPs) belonging to classical depolymerases ($n = 6$), SGNH hydrolases which deacetylate polysaccharides ($n = 6$), and structurally novel RBPs ($n = 2$). Overall, we predicted K-locus specificity for 26 putative depolymerases, of which 12 were deemed as strong predictions against 10 K-loci. In parallel, we attempted recombinant production of 50 putative depolymerases selected from 469 candidate proteins identified in prophages from a representative subset of 99 bacterial isolates, together with an additional 10 depolymerases selected based on GWAS predictions. All recombinant proteins were tested against

a *Klebsiella* reference panel of 119 K-types. Of the 50 manually chosen putative depolymerases, 34 failed to yield detectable recombinant expression, a pattern unlikely to be explained by degraded or defective prophages. Of the 14 active enzymes, 5 targeted a K-locus different from that of their bacterial host, and enzyme specificity could not always be reliably inferred from sequence similarity or structural homology. Comparison of GWAS predictions with experimental validation results revealed that 10 of the 12 strongest GWAS predictors were confirmed experimentally, while 2 produced soluble protein but showed no detectable activity against the tested K-types. Together, these results highlight the intrinsic difficulty of predicting activity and capsule specificity of prophage-encoded RBPs from genomic information alone. Finally, analysis of 4,598 high-completeness prophages revealed that SGNH-domain hydrolases are among the most prevalent enzymatic domains in prophage RBPs. Two SGNH-domain RBPs identified by GWAS were experimentally confirmed as active esterases, supporting capsule deacetylation as a widespread alternative to polysaccharide depolymerisation in temperate phages. Our findings reveal that *Klebsiella* prophages encode structurally diverse RBPs, suggesting temperate phages may rely not only on depolymerisation but also on capsule modification—such as deacetylation—for infection. This also implies that capsule diversity in *K. pneumoniae* may be substantially underestimated, with implications for phage specificity, competition and vaccine design.

Introduction

Host range is among the most critical traits of viruses, determining the variety of hosts that can be infected by a given virus[1]. Understanding and predicting viral host range is fundamental for studying viral evolution, ecology and infection dynamics[2]. In bacteriophages, host range is driven by co-evolution with bacterial hosts[3,4] and shaped by horizontal gene transfer[5,6], with broad relevance for phage therapy and microbiome engineering[7,8,9,10]. However, accurate prediction of viral host range remains challenging due to the complexity of virus-host interactions, from initial adsorption to successful evasion of host defence systems[11]. Nevertheless, recent advances in multi-omics and computational biology have enabled statistical predictions that can be validated empirically[12]. Indeed, two recent genomic studies demonstrate that such approaches can help successfully predict the infectivity of virulent phages in two species of *Enterobacteriaceae*, *Klebsiella pneumoniae*[13] and *Escherichia coli*[14], underscoring the critical role of receptor-binding proteins (RBPs) in virus host range.

K. pneumoniae in particular serves as a powerful system for studying bacteria-virus interactions. It is an opportunistic bacterial pathogen characterised by the extensive diversity of its capsular polysaccharides (CPS)[15]. This diversity encompasses 77 serologically distinct CPS polysaccharides (K-types) with fully characterised biosynthetic loci (K-loci), and at least 86 further genetically distinct K-loci of unknown polysaccharide structure[16,17,18]. Experimentally isolated virulent phages infecting *K. pneumoniae* typically produce visible halos on bacterial lawns, indicating enzymatic degradation of CPS[19]. These halos result from phage RBPs with depolymerase activity, enabling recognition, degradation, and subsequent infection[20]. Such enzymes generally exhibit highly specific activity against individual CPS types, with 58 experimentally validated depolymerases described so far, targeting 31 distinct K-types[21]. Structurally, these depolymerase-containing RBPs are typically homotrimers, featuring a characteristic right-handed parallel β-helix[22,23,24,25]. However, our understanding of the genetic, structural diversity and specificity of depolymerase-containing RBPs in *Klebsiella* phages remains far from complete.

Given that our knowledge about phage depolymerases comes predominantly from studies focused on virulent phages, their role in the biology of temperate phages remains even less understood. In *K. pneumoniae*, studies have demonstrated capsule-specific tropism in temperate phages, observing that prophages induced from bacterial genomes could infect only bacterial lineages with identical K-loci, evidenced by characteristic halos indicative of depolymerase activity[26]. Yet, attempts to computationally

identify corresponding depolymerase genes in prophages failed to show clear correlations with the observed infection patterns. Subsequent genomic analyses of nearly 4,000 *K. pneumoniae* genomes have similarly reported that, while serotype specificity predicts phage-mediated gene flow between bacterial lineages, identifiable depolymerase sequences in prophages remain surprisingly scarce[27]. More recently, a machine-learning study showed that genetic features of prophage-encoded depolymerases could partially predict both the K-locus of the bacterial host and the host range of virulent phages[28]. However, the predictive power of these models was generally low, and depolymerases were detected in only about one-fifth of prophages, raising the question of whether this apparent scarcity reflects technical limitations in annotation or fundamental biological differences between virulent and temperate phages.

In light of these observations, several key gaps remain in our understanding of capsule specificity in temperate *Klebsiella* phages. First, we still do not know which prophage genes determine capsule tropism, or whether these determinants correspond to the canonical depolymerases that mediate host range in virulent phages. Second, depolymerase genes appear strikingly rare in prophage genomes, and it is unclear whether this scarcity reflects a genuine biological difference—such as alternative infection strategies—or simply limitations in our ability to functionally annotate divergent receptor-binding enzymes. Third, even when putative depolymerases can be identified, it remains unknown whether they are active against the capsule type of their host, or any other K-type, given that capsule switching provides a common route for escaping phage infection. Addressing these questions is essential to understand how temperate phages interact with capsulated bacteria and to predict the host range encoded within prophage genomes.

To address these gaps, we aimed to identify molecular determinants of capsule specificity in temperate phages of *K. pneumoniae* using a Genome-Wide Association Study (GWAS) approach. GWAS identifies genetic variants statistically associated with phenotypes of interest, an approach already successfully adapted to bacterial genomics[29,30,31,32]. Compared to other statistical learning methods, GWAS has the advantage of not relying on pre-labelled training datasets, thereby avoiding predefined functional expectations and allowing for biologically interpretable associations between genetic variants and phenotypic traits. We applied GWAS to a diverse dataset of *K. pneumoniae* genomes from two systematically curated collections (Australia[33] and Italy[34]), representing near-random sampling from a collection of 3,900 genomes across diverse ecological and clinical niches. Within this dataset, we predicted prophages and employed GWAS on the subset of 2,527 genomes for which we could confidently predict the type of capsular polysaccharide (K-type), identifying prophage protein clusters statistically associated with

35 common K-types as candidate predictors of capsule specificity. Independently, we manually curated depolymerase-containing RBPs from prophages found in 99 representative bacterial genomes (KASPAH-REF) and performed recombinant protein preparation and enzymatic activity testing against 119 K-type strain collection, using 58 previously characterised depolymerases from virulent phages as a comparative reference[21] (see S1 Table; see also Figure 1 for a visual summary of our approach). Our results demonstrate that while prophages can be a rich source of structurally and functionally diverse depolymerases, they are also an unpredictable one. Our approach also uncovers a previously underappreciated class of receptor-binding proteins dominated by SGNH-domain enzymes, likely involved in removing capsule modifications such as acetyl groups, and pointing to capsule deacetylation as a widespread alternative mechanism of infection initiation.

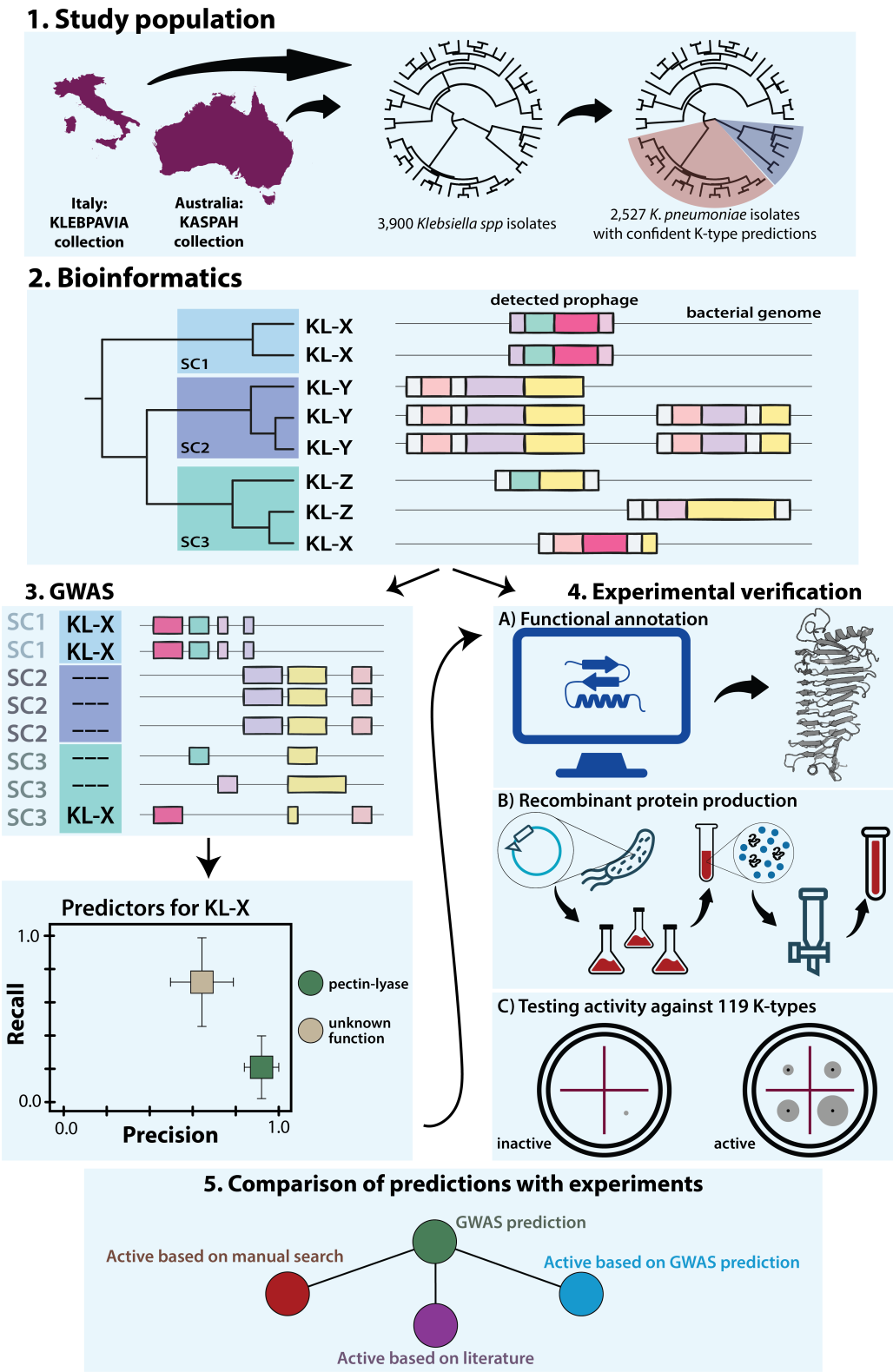
1. Study population
Italy: KLEBPAVIA collection
Australia: KASPAH collection
3,900 Klebsiella spp isolates
2,527 K. pneumoniae isolates with confident K-type predictions
2. Bioinformatics
detected prophage
bacterial genome
SC1
SC2
SC3
KL-X
KL-X
KL-Y
KL-Y
KL-Y
KL-Z
KL-Z
KL-X
3. GWAS
SC1 KL-X
SC1 KL-X
SC2 ---
SC2 ---
SC2 ---
SC3 ---
SC3 ---
SC3 KL-X
Predictors for KL-X
Recall
Precision
1.0
0.0
0.0
1.0
pectin-lyase
unknown function
4. Experimental verification
A) Functional annotation
B) Recombinant protein production
C) Testing activity against 119 K-types
inactive
active
5. Comparison of predictions with experiments
GWAS prediction
Active based on manual search
Active based on GWAS prediction
Active based on literature

Figure 1. Visual summary of the study approach. (1) We assembled an ecologically diverse collection of 3,900 *Klebsiella* genomes, of which 2,527 *Klebsiella pneumoniae* isolates had high-confidence capsular (K-locus) assignments. (2) Prophage regions were predicted in each genome, and all prophage-encoded proteins were clustered and functionally annotated. (3) A bacterial GWAS (elastic-net model) was then performed on 2,527 isolates to link prophage protein clusters to 35 most diverse K-loci, yielding protein clusters as candidate predictors of capsule specificity. (4) 50 manually selected putative depolymerases from 99 high-quality reference genomes ("KASPAH-REF") were selected for recombinant protein production and enzymatic activity testing against a panel of 119 *Klebsiella* strains with diverse K-types. (5) GWAS predictors (green node) were compared to these experimental results (red node) and to previously characterized virulent-phage depolymerases (purple node); for predictors lacking any known sequence similarity, we selected representative β-helix (pectin-lyase domain) proteins for recombinant production and functional testing (blue node). Country outlines were obtained from Natural Earth (Admin 0 – Countries, 1:10m scale; `https://www.naturalearthdata.com/downloads/10m-cultural-vectors/`).

Results

Isolate selection and data

The aim of this study was threefold: (a) to identify genetic determinants of capsule tropism in temperate phages of *K. pneumoniae*, (b) to functionally characterise putative depolymerase candidates from a subset of isolates, and (c) to compare genomic predictions with experimental outcomes. To this end, we carried out a comprehensive search that is described in detail in Methods and visually summarised in Figure 1. Briefly, we assembled a dataset of 3,900 *Klebsiella* genomes by combining two large and systematically curated sources. The KASPAH dataset includes 391 clinical and colonisation (gut/throat) isolates collected from hospital patients in Melbourne[35,36,33], while KLEBPAVIA comprises 3,509 isolates sampled through a One Health framework in northern Italy, spanning human, animal, environmental, and agricultural sources[34]. Together, these datasets provide a comprehensive and ecologically broad view of *Klebsiella* diversity, avoiding the clone- and outbreak-driven biases often seen in public genome databases. All isolates were sequenced using a short-read technology, and 99 representative KASPAH genomes were additionally sequenced using long reads to produce high-quality hybrid assemblies[37] (KASPAH-REF). For each genome, we identified the capsular locus (K-locus) using Kaptive[17] and assigned bacterial isolates into sequence clusters (SCs) using PopPunk[38]; the latter was done to account for population structure in the GWAS analysis, minimising confounding from recent shared ancestry and prophage co-inheritance among closely related isolates.

Prophage regions ($n = 8105$) were identified using VirSorter[39], PhiSpy[40], and CheckV[41], retaining both high-completeness (likely intact) and medium-completeness predictions to capture fragmented prophages in short-read assemblies (see S1 Figure for a distribution of prophage completeness in most frequent K-types). To facilitate robust comparisons, we focused subsequent analyses on 2,527 genomes within the *K. pneumoniae* species complex (KPSC) that had confident K-locus predictions, and focused on 35 K-loci with sufficient diversity ($\geqslant$ 10 SCs) to ensure statistical power in subsequent analyses. This dataset is visualised in Figure 2, with the bacterial phylogeny shown in Fig. 2A and bacterial and prophage diversity across K-loci shown in Fig. 2B. The 35 K-loci each represented between 10-24 SCs and from 29 phage variants for KL47 to 112 phage variants for KL30 (median of 55), hence encompassing considerable diversity in both capsule type and prophage content. The full distribution of SCs and phage variants per K-locus is shown in S2 Figure. The bacterial and prophage metadata are provided in S2 and

S3 Tables, respectively.

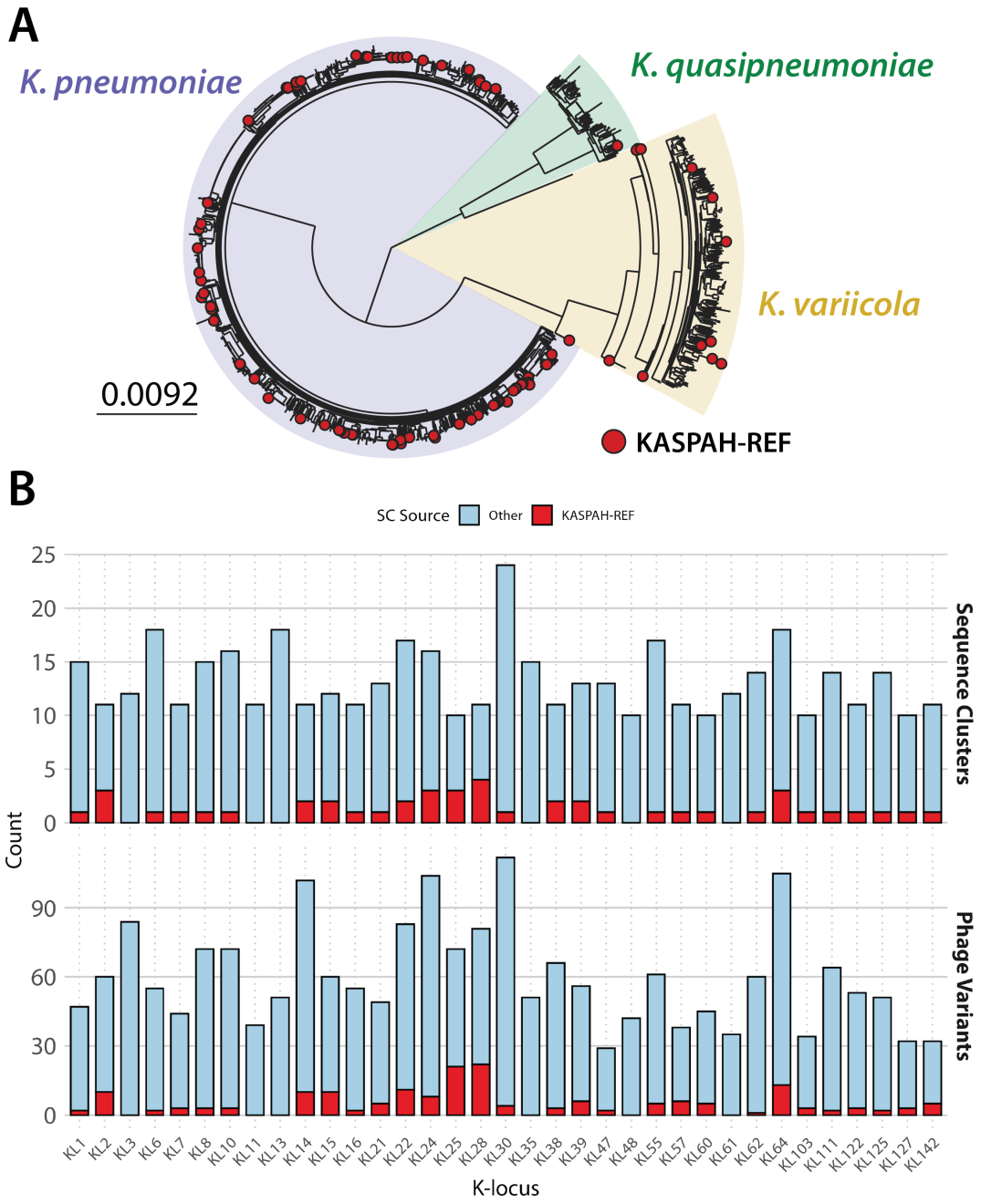

Figure 2. The dataset. (A) Maximum-likelihood tree of 2,527 isolates from the *K. pneumoniae* Species Complex (KPSC). Red dots show isolates belonging to the KASPAH-REF subset ($n = 99$). Shaded areas show the three main species from the KPSC with reliable K-locus predictions: *K. pneumoniae* (purple), *K. variicola* (yellow), and *K. quasipneumoniae* (green). (B) Top plot shows bacterial diversity as the number of sequence clusters (SCs) per predicted K-locus amongst KASPAH-REF isolates (red) and all other isolates (blue). Bottom plot shows phage diversity measured by the number of phage variants (similar prophages clustered using the wGRR = 0.95 threshold; see Methods) identified in the KASPAH-REF isolates (red) and using all 2,428 remaining isolates (blue). The plot includes only the 35 most diverse K-loci in our dataset, defined as those present in at least 10 different SCs; these 35 K-loci were selected for GWAS analyses.

GWAS predicts structurally diverse receptor-binding proteins

To perform a genome-wide association study (GWAS), we analysed prophage protein clusters as genetic predictors of capsule tropism by correlating their presence with K-locus assignments. We used an elastic-net regression model— a machine-learning extension of linear regression designed for genomic settings with many correlated predictors. In our case, the predictors are the presence or absence of each prophage protein cluster in an isolate, and the response variable is the presence or absence of a given K-locus. Because multiple protein clusters can co-occur within prophages and many are homologous, this results in a high-dimensional predictor matrix with substantial correlation structure. Elastic-net accommodates this by combining lasso, which selects a small number of informative clusters by shrinking others to zero, with ridge regression, which retains groups of correlated predictors by shrinking their coefficients together. This allows the model to detect a single dominant capsule-specific RBP when one exists, while also capturing cases where several related or mosaic proteins correlate with the same K-type – a behaviour that matches the biological diversity of phage RBPs. To minimise confounding from lineage effects, population structure was incorporated by including sequence clusters (SCs) as covariates in the regression model. The elastic-net procedure then estimates a regression coefficient (β) for each protein cluster, indicating the strength and direction of its association with the given K-locus. Predictors were retained when $\beta > 0$ and the p-value for that coefficient was below 0.05 (following Bonferroni-correction for multiple testing).

To evaluate the performance of elastic-net predictors, we used standard classification metrics based on true and false predictions. In our setting, a true positive (TP) is a protein cluster present in an isolate with the corresponding K-locus, a false positive (FP) is a protein cluster present in an isolate with a different K-locus, and a false negative (FN) is a protein cluster absent from an isolate with the corresponding K-locus. Precision ($\mathrm{TP}/(\mathrm{TP}+\mathrm{FP})$) measures the proportion of predictor occurrences that correctly identify their associated K-locus, whereas recall ($\mathrm{TP}/(\mathrm{TP}+\mathrm{FN})$) measures the proportion of isolates with a given K-locus in which the corresponding predictor is detected. The F1 score combines precision and recall into a single harmonic mean, favouring predictors that balance both metrics. The Matthews Correlation Coefficient (MCC) incorporates all four quantities (TP, FP, FN, true negatives), providing a more robust measure of overall classification performance, especially for imbalanced datasets such as ours. Full details of the GWAS methodology are provided in Methods.

Figure 3A shows the results of GWAS for predictors filtered with F1 $\geqslant 0.5$ and MCC $\geqslant 0.5$. Using

those thresholds, we obtained 35 protein clusters associated with 24 out of 35 K-loci, with eight K-loci (KL3, KL11, KL14, KL21, KL24, KL25, KL111 and KL127) having multiple associated predictors. Functional annotation of these 35 clusters using PHROGs[42] revealed that 23 of them encoded tail-associated proteins, of which 18 were predicted to be tail fibres—consistent with our expectation that capsule-specific associations would predominantly involve receptor-binding proteins (RBPs). We also observed a negative correlation between precision and recall across predictors (Pearson's $r = -0.58$, 95% CI: -0.76 to -0.30; see S3 Figure). While such a trade-off is expected, the presence of low-precision, high-recall predictors suggests that some candidates may target multiple K-types or represent associations unrelated to capsule interaction (more on this later).

Given that the initial GWAS results predominantly identified receptor-binding proteins (RBPs), we next focused on refining predictors based on biological expectations of polysaccharide specificity. Since most known depolymerases act on a single capsule type[21], we applied a stringent precision threshold ($\geqslant 0.8$) to prioritise highly specific predictors. Furthermore, since RBPs are often modular, with variation across N-terminal and C-terminal domains[43], we reasoned that the clustering thresholds used to define protein families could affect whether related proteins were grouped into single predictors or split into distinct predictors. To address this, we systematically repeated GWAS across six clustering thresholds, combining identity cutoffs (none, 50%, 80%) with coverage thresholds (50%, 80%), maximising chances that such potential modularity was captured. For each K-locus, we then selected the clustering level that maximised predictor performance (defined as the highest product of F1 score and MCC, with both metrics $\geqslant 0.5$), aiming to identify the most biologically informative protein cluster for each of them.

Using these criteria, we obtained predictions for 16 out of 35 K-loci (Figure 3B). We then selected a representative protein from the best predictor (protein cluster) and generated its 3D models using AlphaFold3[44]. (Note that for 19 K-loci, there were no predictors that fulfilled our criteria; see Discussion.) We also mapped the locations of the major ECODs (Evolutionary Classification of Protein Domains) detected in those proteins onto the 3D structures (see Figure 3B). We found that 6 of those 16 proteins (from predictors for KL3, KL14, KL28, KL38, KL62, KL127) were classical depolymerases with the characteristic, parallel β-helix fold (ECOD T-level: 207.2.1, *Pectin-lyase like*). Another 6 proteins (from predictors for KL2, KL8, KL11, KL16, KL24 and KL55) lacked the parallel β-helix, carrying a *SGNH hydrolase* domain (ECOD T-level: 2007.5.1), and four remaining proteins had no significant ECOD hits identified. Of those four, two proteins (predictors for KL47 and KL103) looked like unusual RBPs with

previously uncharacterised folds, while the remaining two (predictors for KL30 and KL111) did not look like RBPs. We used FoldSeek[45] to search the non-RBP predictors against the CATH Protein Structure Classification database[46], the Protein Data Bank database[47], and the AlphaFold Protein Structure database[48] and found that these proteins structurally resembled known acetyltransferases.

Finally, we examined how varying threshold for filtering GWAS predictors influenced the distribution of ECOD domains identified among the associated protein clusters. Regardless of the filters used, filtered predictors were dominated by two structural folds: the *Pectin lyase-like* domain and the *SGNH hydrolase* domain, along with a substantial fraction of proteins lacking functional annotation (Figure 3C). These results suggest that capsule tropism in temperate phages of *K. pneumoniae* involves not only classical depolymerases but also other structurally diverse RBPs with distinct folds. Furthermore, we observed that by relaxing predictor filtering criteria — specifically, allowing lower recall while maintaining high precision — additional lower-frequency protein clusters emerged, some of which may encode previously uncharacterised β-helix-containing depolymerases. We therefore next lowered the filtering criteria and applied a series of manual curation steps to identify putative depolymerases against as many K-loci as possible.

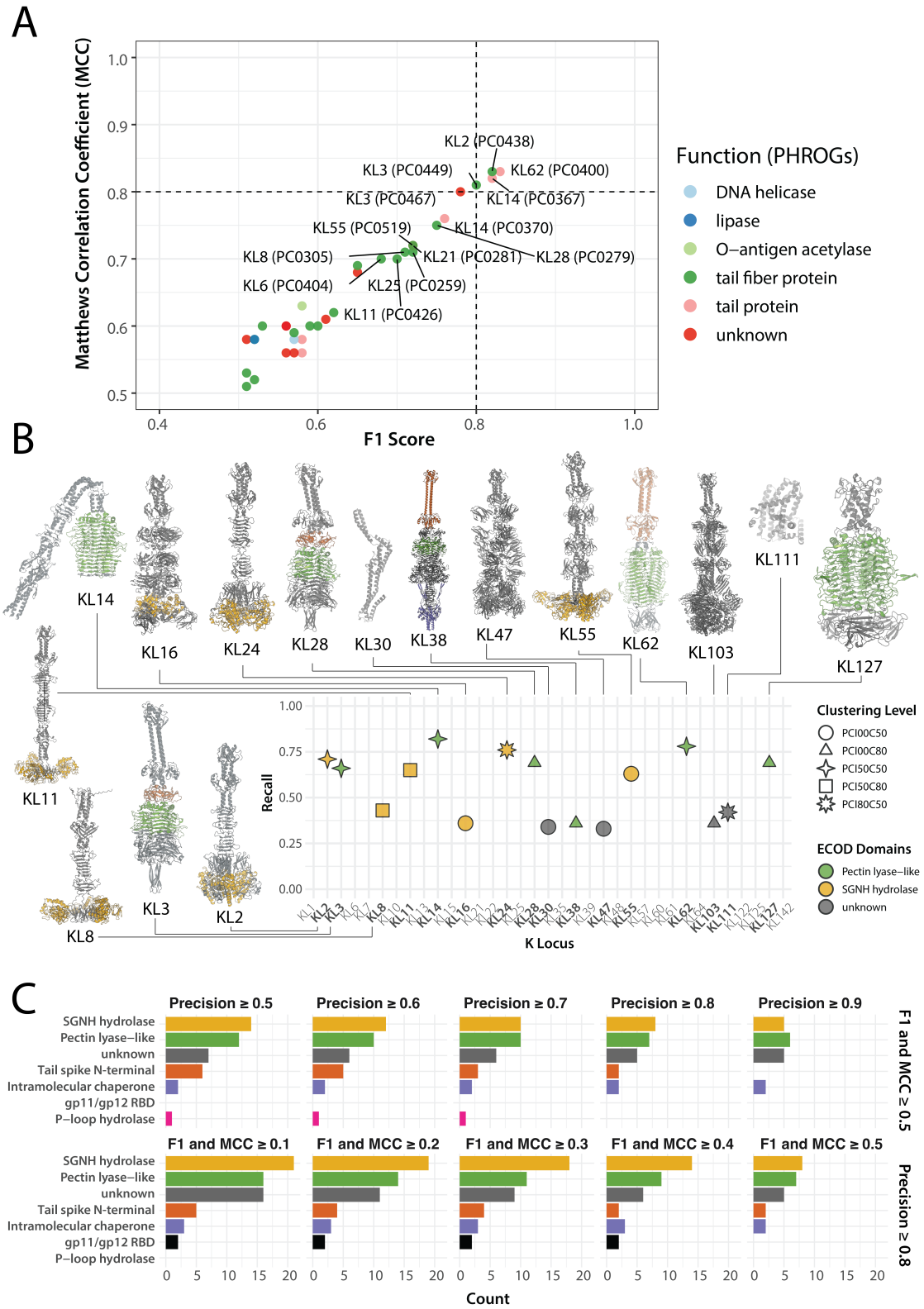
A
Matthews Correlation Coefficient (MCC)
F1 Score
KL2 (PC0438)
KL3 (PC0449)
KL62 (PC0400)
KL3 (PC0467)
KL14 (PC0367)
KL55 (PC0519)
KL14 (PC0370)
KL8 (PC0305)
KL21 (PC0281)
KL28 (PC0279)
KL6 (PC0404)
KL25 (PC0259)
KL11 (PC0426)
Function (PHROGs)
DNA helicase
lipase
O–antigen acetylase
tail fiber protein
tail protein
unknown
B
KL14
KL16
KL24
KL28
KL30
KL38
KL47
KL55
KL62
KL103
KL111
KL127
KL11
KL3
KL2
KL8
Recall
K Locus
Clustering Level
PCI00C50
PCI00C80
PCI50C50
PCI50C80
PCI80C50
ECOD Domains
Pectin lyase–like
SGNH hydrolase
unknown
C
Precision ≥ 0.5
Precision ≥ 0.6
Precision ≥ 0.7
Precision ≥ 0.8
Precision ≥ 0.9
F1 and MCC ≥ 0.5
F1 and MCC ≥ 0.1
F1 and MCC ≥ 0.2
F1 and MCC ≥ 0.3
F1 and MCC ≥ 0.4
F1 and MCC ≥ 0.5
Precision ≥ 0.8
SGNH hydrolase
Pectin lyase–like
unknown
Tail spike N-terminal
Intramolecular chaperone
gp11/gp12 RBD
P–loop hydrolase
Count

Figure 3. GWAS predicts diverse receptor-binding proteins. (A) F1 score versus Matthews Correlation Coefficient (MCC) for prophage protein clusters (PC50 clustering) identified as GWAS predictors with F1 $\geqslant$ 0.5 and MCC $\geqslant$ 0.5. Colours indicate functional annotations assigned using PHROGs. (B) Recall of the strongest predictor identified for each K-locus. For each K-locus, predictors with precision $\geqslant$ 0.8 were selected, retaining the one with the highest F1$\times$MCC product across clustering thresholds; only those with F1 $\geqslant$ 0.5 and MCC $\geqslant$ 0.5 are shown (with the corresponding K-loci names bolded). Shapes indicate the clustering threshold at which the strongest predictor was identified; colours reflect the top predicted ECOD domain, or grey for unannotated regions. For the 16 predictors obtained, AlphaFold3 models were generated based on representative proteins, with ECOD domains coloured based on HHpred predictions. (C) Distribution of ECOD domains across GWAS predictors across different filtering thresholds, shown as bar plots. The top row summarises predictors meeting F1 $\geqslant$ 0.5 and MCC $\geqslant$ 0.5, across varying precision thresholds (0.5–0.9). The bottom row shows predictors filtered at minimum precision $\geqslant$ 0.8, across varying F1 and MCC thresholds (0.1–0.5). ECOD domain names are abbreviated as follows: *SGNH hydrolase* (yellow; T-level: 2007.5.1), *Pectin-lyase like* (green; T-level: 207.2.1), Tail spike N-terminal (brown; Full name: *Putative tailspike protein Orf210 N-terminal domain*, T-level: 3856.1.1), Intramolecular chaperone (purple; Full name: *Intramolecular chaperone domain in virus tail spike protein*, T-level: 3240.1.1), gp11/gp12 RBD (black; Full name: *gp11/gp12 receptor-binding domain*, T-level: 877.1.1), P-loop hydrolase (pink; Full name: *P-loop containing nucleoside triphosphate hydrolases*, T-level: 2004.1.1). Grey denotes the absence of functional hits.

Prophages are an unpredictable source of active depolymerases

To comprehensively identify GWAS-predicted depolymerases containing the characteristic right-handed β-helix fold, we developed a semi-supervised curation framework combining multiple evaluation steps (full details in S1 Text). Starting from statistically significant GWAS predictors, we first assessed precision and recall for each protein cluster at both isolate and sequence-cluster (SC) levels to account for population structure. To accommodate the extensive modularity of phage receptor-binding proteins (RBPs), we then evaluated six alternative clustering thresholds (coverage thresholds of 50% and 80%; sequence identity thresholds of 0%, 50%, and 80%) to identify parameter settings that best captured biologically coherent protein families for each K-locus. Next, we applied phylogenetic mapping using Phandango[49] to visualise whether predictors consistently tracked with K-locus distribution across the bacterial tree. Finally, based on combined evidence from classification metrics, clustering behaviour, domain annotation and structural modeling, predictors were conservatively classified as 'strong' or 'likely'. Using this approach, we identified 12 'strong' predictors for 10 K-loci (KL3, KL10, KL14, KL25, KL28, KL38, KL60, KL62, KL64, KL111) and 14 'likely' predictors for 11 K-loci (KL7, KL14, KL15, KL22, KL24, KL25, KL30, KL47, KL64, KL122, KL127), together spanning 18 of the 35 analysed K-loci (Figure 4A).

Separately, we experimentally screened prophage-encoded candidate depolymerases from the KASPAH-REF dataset, as explained in detail in the Methods. Briefly, starting from 469 candidate proteins meeting initial selection criteria (minimum length of 500 aa, functional annotation as tail fibre or tail spike, predicted presence of a parallel β-helix domain) , we clustered these into 124 sequence groups, modelled all representatives using AlphaFold2[50], and manually evaluated the resulting structures for conserved features characteristic of depolymerases, particularly the parallel β-helix fold. This process yielded 50 candidate sequences from 32 clusters for recombinant protein expression (see Methods). For each of these 50 candidates, we attempted recombinant production and purification, followed by enzymatic testing against 119 diverse *Klebsiella* strains (S4 Table). Out of the 50 candidates, 34 recombinant proteins received insufficient yield, 2 were produced as soluble proteins but showed no detectable activity on any of the 119 K-types, and 14 yielded soluble, active enzymes targeting one of 11 K-types (KL23, KL28, KL32, KL38, KL46, KL52, KL60, KL62, KL64, KL127 and KL143; see Figure 4A, S5A and S5B Table and S2 Text).

Given that 34 of the 50 recombinant expression attempts resulted in low or undetectable yields and no measurable activity, we next examined whether insufficient expression could account for the apparent

lack of function. Notably, several of the depolymerases that we confirmed as active were also expressed at low levels, yet even small amounts of protein were sufficient to degrade capsule material (see S5A Table). These conditions were consistent with, and in some cases directly matched, those reported in the literature for successful expression of large *Klebsiella* phage depolymerases (see S5B Table). To assess whether expression constraints were responsible for the inactivity of the remaining candidates, we undertook a systematic re-expression campaign in which eight cluster representative proteins were expressed under a new set of conditions (IPTG inducer concentration, expression temperature, density of the culture induced, and expression duration). Despite these modifications, seven of the eight candidates showed no detectable overexpression by SDS–PAGE, and the single protein that did express at low levels remained inactive against all tested *Klebsiella* isolates (S5C Table).

Structural modelling of non-expressed proteins revealed no obvious folding defects, with predicted architectures indistinguishable from active homologues (S5 and S6 Figures). Moreover, there was no evidence that prophages from which low-yield proteins originated were more likely to be cryptic (S7 Figure) and, in several cases, co-encoded both active and inactive depolymerases (S8 Figure). We also found two pairs of highly similar proteins (amino acid identity $> 97\%$, but modification accumulation in different enzyme domains), where one was recombinantly obtained as soluble and active, and the other was not expressed (S9 Figure). These findings suggest that the failure to obtain soluble, active enzymes is unlikely to be attributable to cryptic prophages or obvious structural degradation, but rather reflects the difficulty of heterologously expressing large, trimeric proteins outside their native phage context.

The 26 GWAS predictors (from both the 'strong' and 'likely' categories) were first compared to the manually curated collection of 50 candidate depolymerases from the KASPAH-REF isolates (S5A Table) based on their protein sequence similarity. When the GWAS predictor from the 'strong' category did not not display sequence similarity to any protein within our recombinant-protein set, or any of the available enzymes with confirmed specificity described already in the literature (S1 Table), we commissioned synthetic plasmid preparation and recombinant protein production for one representative protein from each cluster based on manual analyses to maximise chances of its successful production and enzymatic activity (see Methods).

Figure 4 shows a direct comparison of computational and experimental results per K-locus, alongside additional recombinant proteins from the literature from virulent phages. Specifically, 4A shows a table of the number of predictors for each K-locus, and Figure 4B shows a network-like comparison of the two

sets of proteins, alongside 58 previously published depolymerases from virulent phages[21] (see S1 Table). We found that, out of 12 predictors in the 'strong' category, 5 were confirmed by sequence or structural similarity to the recombinant protein analysis (KL28, KL38, KL60, KL62, KL64) and 2 were confirmed by sequence or structural similarity to recombinant proteins from the literature: KL3 had distant similarity and a visibly different C-terminus, while the KL25 prediction had structurally identical receptor-binding domain (see S3 Text). Of the remaining 5 predictors, which were used to select 5 proteins for commercial production, 3 were confirmed as active enzymes (KL3, KL62 and KL111) while 2 did yield soluble proteins (KL10 and KL14) but no activity was detected against any of the 119 K-types. Additionally, one protein commercially produced based on a 'likely' predictor (KL122) was found active against isolates with the KL122 locus (Figure 4B; S5A Table).

By contrast, out of 11 K-types for which we obtained active enzymes via the manual search in the KASPAH-REF collection, five specificities (K23, K32, K46, K52, KL143) could not be compared with GWAS results, as these K-types were excluded from the set of 35 loci, hence no predictors were generated. (For one of them, KL23, a virulent-phage protein had previously been reported in the literature and we observed structural homology of the receptor-binding domain; see S3 Text.) The enzymatic activity of the remaining six KASPAH-REF K-specific enzymes (K28, K38, K60, K62, K64 and KL127) was in line with GWAS predictions, either predictors classified as 'strong' or 'likely'. Importantly, 5 of 14 enzymes were active against a different K-type than the K-type of the prophage host (see S4 Figure).

In summary, we found that (a) of the 12 'strong' GWAS predictors, 10 were experimentally validated and 2 remained inconclusive, and (b) among the 14 active enzymes identified through manual search, only 6 corresponded to the 35 K-loci included in the GWAS analysis; all 6 were supported by GWAS predictions ('strong' or 'likely' predictors). Importantly, prophages proved to be an unpredictable source of active enzymes as most candidate proteins resulted in a low yield, and several enzymes displayed activity against a different K-type than that of their host strain (see also Discussion).

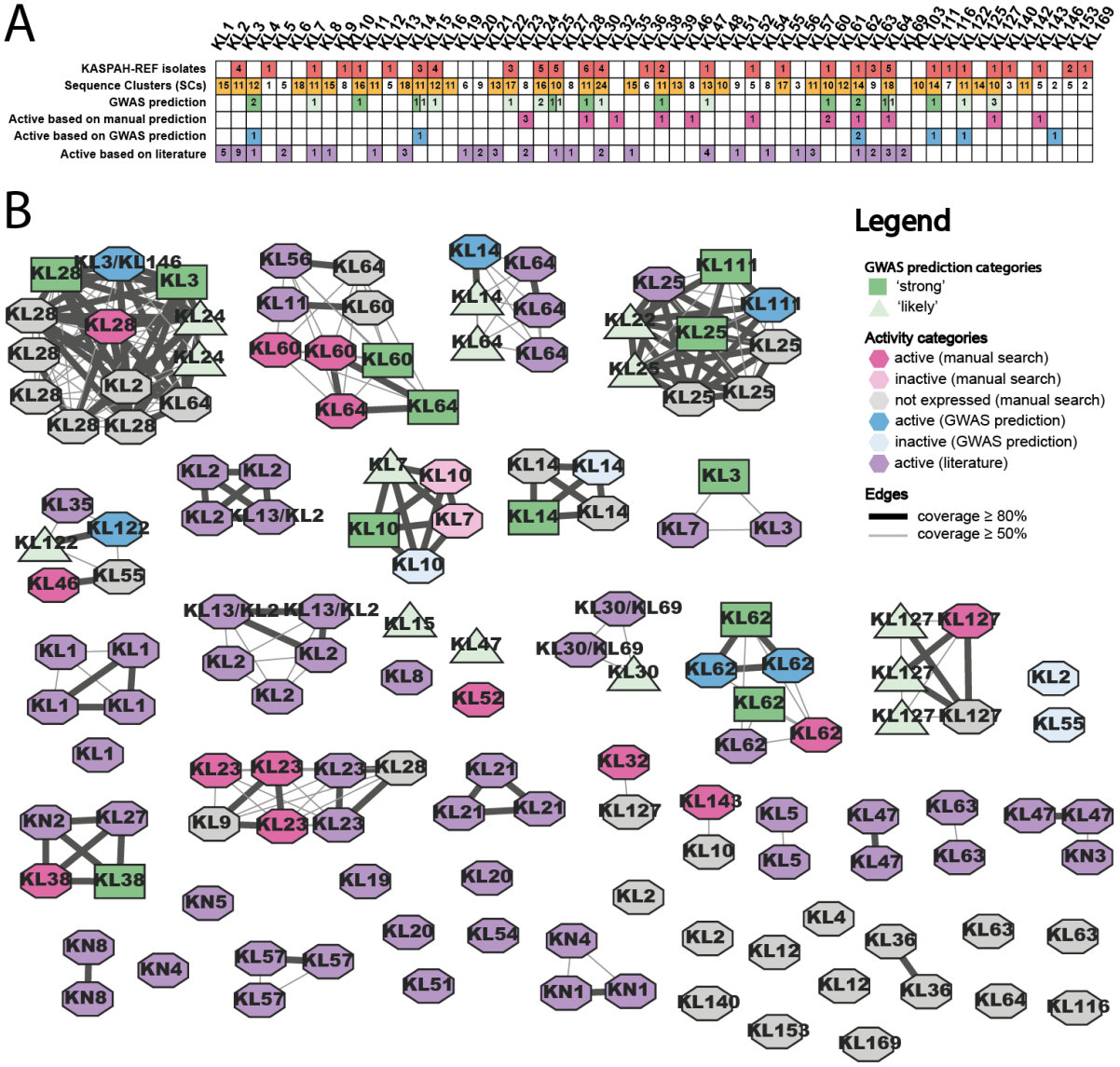

Figure 4. Prophages can be an unpredictable source of active depolymerases. (A) Table shows a systematic comparison of the number of bacterial SCs from the KASPAH-REF collection (red), the number of bacterial SCs from the full dataset (yellow), number of GWAS predictions in the 'strong' (dark green) 'likely' (light green) category), number of active recombinant proteins based on manual (magenta) and based on GWAS' 'strong' predictors (dark blue), and the number of experimentally verified depolymerases from virulent phages from Cheetham and colleagues[21]. (B) Sequence similarity network where each node represents a protein sequence and edges link proteins with a minimum of 50% bidirectional coverage (thin edges) or a minimum of 80% bidirectional coverage (thick edges). Colours are in accordance with panel A, with blue nodes showing additional proteins obtained via gene synthesis and recombinant protein production.

Complexity of the genotype–phenotype relationships in phage depolymerases

While the overall agreement between GWAS predictions and experimental validation was strong, closer inspection of the sequence similarity network (Figure 4B) revealed multiple genetic links between enzymes with different K-type specificities (confirmed or predicted). To explore this complexity more systematically, we performed a detailed comparison of the protein sequence, structure and specificity relationships across these enzymes, as described in S3 Text. This analysis revealed several key patterns that exemplify the challenges in predicting capsule specificity from sequence – or even structure – alone.

First, we observed frequent modularity, particularly variation in N-terminal domains known to be related to the virion taxonomic or morphological groups and to vary among phage depolymerases[43] (see S3 Text). Second, we found that enzymes that shared sequence and structural similarity often exhibited distinct substrate specificities, indicating that proteins with related structures of their receptor-binding domains might be functionally divergent. Figure 5 demonstrates this using the example of enzymes known or predicted to target K-types KL3, KL24 and KL28 from two connected components of the network in Figure 4B. We found that GWAS predictors for KL3 (PC0449) and KL28 (PC0279) were both confirmed experimentally (391_03 and 1251_37); a comparison revealed that they are genetically and structurally related to each other and to the KL24 predictions (PC0406 and PC1397; Figure 5A). Nevertheless, recombinant depolymerase against K3 (391_03) and against K28 (1251_37) did not cross-react against K3/K24/K28 reference strains (Figure 5B) even though the K3, K24 and K28 polysaccharide structures are biochemically related (Figure 5C). Furthermore, comparison to another recombinantly tested depolymerase (NCBI accession YP_03347555.1) from a virulent phage active on a K3-type capsule showed no genetic or structural similarity to any of the other ones.

Interestingly, we found that the KL3 recombinant protein (391_03) was also active against the KL146 strain (see Figure 5B). We hypothesised that the dual specificity of the 391_03 protein against KL3 and KL146 reflects underlying structural similarity between the corresponding polysaccharides. Because the polysaccharide structure of KL146 has not yet been resolved at the time of this publication, we instead compared the genetic composition of the KL3 and KL146 loci. We found that the two loci shared two serotype-specific genes: *wcuC* (glycosyl hydrolase) and *wbaZ* (glycosyl transferase), suggesting the presence of a common biochemical motif that could potentially explain the dual specificity of the KL3 prediction (see S10 Figure). Intriguingly, a recent report showed that *K. pneumoniae* isolates carrying the KL146 locus did not react with any available reference antisera, including those against K3 or other

defined capsular types[51].

The network analysis (Figure 4B) further revealed depolymerases with identical substrate specificity whose structural models indicate homology, even though their primary sequences are markedly divergent. Examples include the K1-specific depolymerases (NCBI accessions BAW85698.1 and YP_009302756.1, protein sequence identity of 35%), K5-specific depolymerases (APZ82848.1 and WNO47121.1, identity of 32%) or K21-specific depolymerases (BAW85693.1 and YP_003347556.1, identity of 41%). Therefore, our results indicate that genetic or structural similarity is often insufficient to infer depolymerase specificity, pointing to a nuanced genotype–phenotype map of these enzymes.

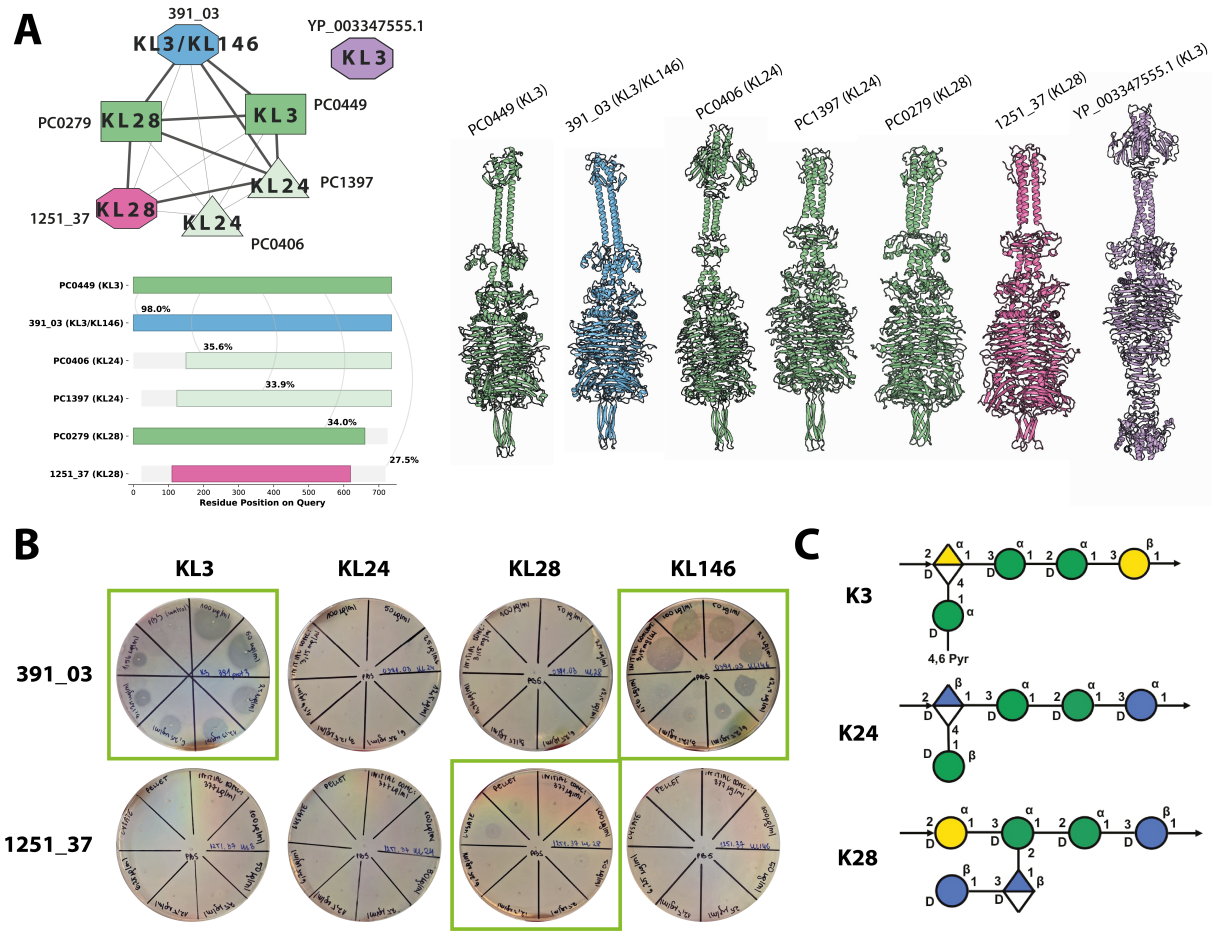

Figure 5. Depolymerase specificity is a complex phenotype. (A) One of the connected components from Figure 4B showing seven depolymerases: two 'strong' GWAS predictions (KL3, KL28; dark green), two 'likely' GWAS predictions (KL24; light green), one active protein from the manual search (KL28; pink), one active protein from GWAS prediction (KL3/KL146; blue), and an active protein from a virulent phage (KL3; purple). All of these proteins, except the virulent KL3 depolymerase, are genetically and structurally related, seen in a systematic visualisation of pairwise alignments and 3D models obtained via AlphaFold3. (B) Spot test assay validating the predicted capsular specificity of 391_03 and 1251_37 selected proteins against reference serotypes KL3, KL24, KL28 and KL146. Plates showing detectable enzymatic activity (halo formation) are highlighted with green boxes. (C) Schematic representations of the K3, K24, and K28 capsule polysaccharides, as obtained from K-PAM[52].

K. pneumoniae prophage RBPs contain diverse enzymes

Our results showed that the β-helix-containing depolymerases are present in prophages, are relatively rare, highly diverse and not always active. Additionally, the 'pectin lyase-like' domain was not the predominant domain detected amongst the top GWAS results. Therefore, we next investigated the prevalence of different protein domains in prophage RBPs from our dataset. To this end, we analysed ECOD domain diversity across 4,598 high-completeness prophages, focusing on prophage protein clusters annotated as RBPs. For each cluster, we classified detected ECOD domains and quantified K-locus diversity via the Simpson index, which reflects the variation of K-loci associated with bacteria in which these proteins were found. We reasoned that capsule-specific RBPs should predominantly occur in genomes sharing the same K-locus.

The association between ECOD domains and K-locus diversity varied substantially across functions. Only four ECOD domains were linked to protein clusters predominantly associated with a single K-locus or few K-loci: *Phage tail fiber protein trimerization domain* (T-level: 79.1.1), *Intramolecular chaperone domain in virus tail spike protein* (T-level: 3240.1.1), *Pectin lyase-like* (T-level: 207.2.1), and *SGNH hydrolase* (T-level: 2007.5.1; see Figure 6A). Among these, only the *Phage tail fiber protein trimerization domain* and the *Pectin lyase-like* domain were predominantly found in clusters with low K-locus diversity, suggesting a stronger association with capsule specificity. In contrast, the *Intramolecular chaperone domain in virus tail spike protein* (T-level: 3240.1.1), *gp11/gp12 receptor-binding domain* (T-level: 877.1.1), and *Uncharacterized protein CV0426* (T-level: 3809.1.1) tended to occur in clusters spanning multiple K-loci, indicating broader distribution. Notably, the *SGNH hydrolase* domain exhibited the greatest variability in K-locus diversity, and on average, was less K-locus-specific than the *Pectin lyase-like* domain.

To investigate how the frequency of RBP domains varies between different K-loci, we mapped the ECOD domains from Figure 6A on the high-quality prophages found in bacterial genomes with various K-loci (see Figure 6B). Most of the RBPs detected carried the 'SGNH hydrolase' domain, while the 'Pectin lyase-like' domain was on average less prevalent, usually co-occurring with other domains (see S11 Figure). Furthermore, we found examples where while no domains were detected in prophage RBPs, adjacent proteins contained hits to 'SGNH hydrolase' (see S12 Figure). Altogether, these results highlight the structural diversity of RBPs found in *K. pneumoniae* prophages and point to the prevalence and importance of alternative capsule recognition mechanisms like SGNH hydrolases in interactions between bacteria and their temperate phages.

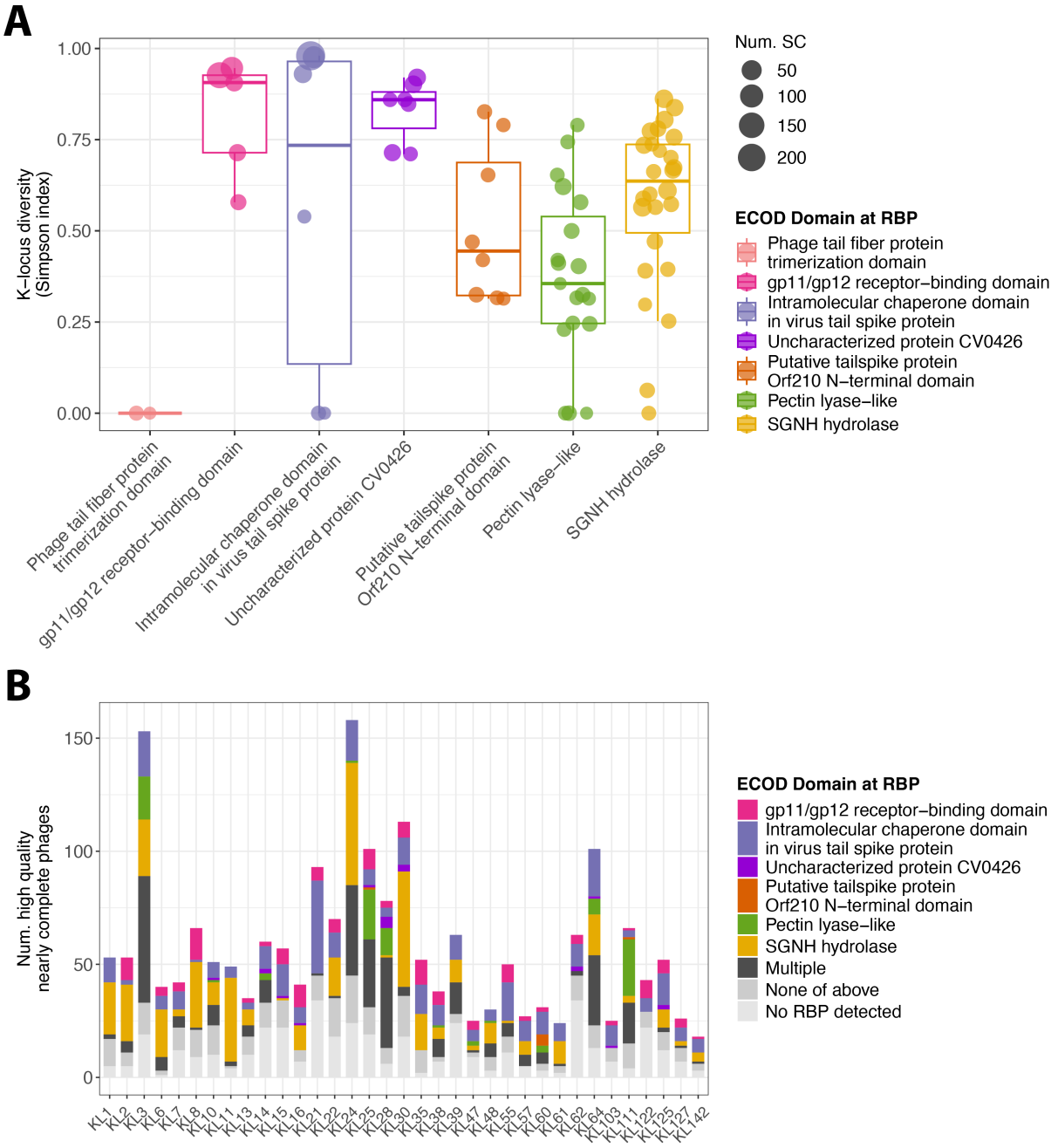

Figure 6. SGNH hydrolases are prevalent amongst *K. pneumoniae* prophage RBPs. (A) Capsule-type (K-locus) diversity within RBP clusters, coloured by their dominant ECOD domain. Prophage proteins from four *Klebsiella* species were clustered at ⩾50% identity and ⩾80% coverage, then RBPs were identified by PHROG annotation (tail fiber or tail spike; probability ⩾90%, qcov ⩾ 50%, scov ⩾ 50%). ECOD domains were assigned to each RBP cluster (probability ⩾ 90%, tcov ⩾ 10%), allowing multiple domains per cluster. Each point shows a cluster with ⩾10 members from ⩾3 SCs, its Simpson index of K-locus diversity (0 = uniform, 1 = maximally diverse), and its ECOD domain. (B) Number of high-quality, nearly complete prophages encoding each ECOD domain within PHROG-defined RBPs, stratified by K-locus (completeness of ⩾90%, n =4,598, of which n =2,047 belonged to the 35 most diverse K-loci). Bars are coloured by the single ECOD domain detected (or light grey if no PHROG RBP was detected, dark grey if multiple ECOD domains co-occur). The most frequent multi-domain combinations include gp11/gp12 + SGNH hydrolase and pectin-lyase-like + Orf210 N-terminal or gp11/gp12 domains (see S11 Figure). Prophages whose RBPs lack any of the domains shown in (A) often encode adjacent proteins with other enzymatic folds (e.g., cysteine proteases, triphosphate hydrolases, SGNH hydrolases; S12 Figure).

Tail fibres with SGNH hydrolase domains are functional esterases

Our analyses revealed that SGNH hydrolases represent one of the most frequent domain families in prophage receptor-binding proteins (RBPs) of *K. pneumoniae*. SGNH-containing RBPs appeared as recurrent GWAS hits across multiple K-loci and, in some K-loci like KL2, KL6 or KL11, nearly all prophages carried a RBP with this domain. A RBP with an SGNH hydrolase domain was previously shown as an active esterase in an *E. coli* phage G7C, which acts as a deacetylase removing an O3-acetyl group from the bacterial lipopolysaccharide[53]. Together, the prevalence of SGNH-domain RBPs across multiple K-loci and their repeated identification as GWAS predictors led us to hypothesise that these prophage-encoded proteins function as active acetyl esterases.

To assess the enzymatic activity of SGNH-domain RBPs in *Klebsiella* prophages, we recombinantly produced two proteins encoded by the strongest GWAS-predicted SGNH-domain candidates: protein 164_08, selected based on the GWAS predictor for KL2 (hereafter 164_08-KL2), and protein 174_38, selected based on the GWAS predictor for KL55 (hereafter 174_38-KL55; cf. Fig. 3). AlphaFold3 modelling showed that both proteins are tail fibres with a characteristic SGNH-containing, brush-like C-terminal architecture (Fig. 7A). As a negative control, we included 184_43, selected based on the GWAS predictor for KL111 (hereafter 184_43-KL111), which had a parallel β-helix domain but lacked an SGNH domain (see Fig. 7A). To functionally characterise these proteins, we employed an established surrogate assay based on hydrolysis of *p*-nitrophenyl acetate (pNPA). SGNH hydrolases – including characterised carbohydrate esterases and polysaccharide deacetylases – readily cleave simple acetate esters such as pNPA, and this reaction is widely used as a diagnostic readout of SGNH catalytic activity independent of the native polysaccharide substrate[54,55,56]. Hydrolysis of pNPA releases *p*-nitrophenol and acetic acid, enabling robust detection through absorbance at 405 nm and pH-dependent colourimetric shifts, respectively.

Both 164_08-KL2 and 174_38-KL55 efficiently hydrolysed pNPA, exhibiting a clear time-dependent increase in absorbance and visible yellowing consistent with *p*-nitrophenol formation (Fig. 7B). A complementary phenol-red pH-shift assay confirmed acid release during the reaction, further supporting ester cleavage (Fig. 7C). In contrast, 184_43-KL111 and all no-substrate controls showed no detectable activity. Together, these assays demonstrate that the SGNH-domain tail fibres encoded by *Klebsiella* prophages function as active acetyl esterases, supporting their proposed role in capsule modification rather than depolymerisation (see Discussion).

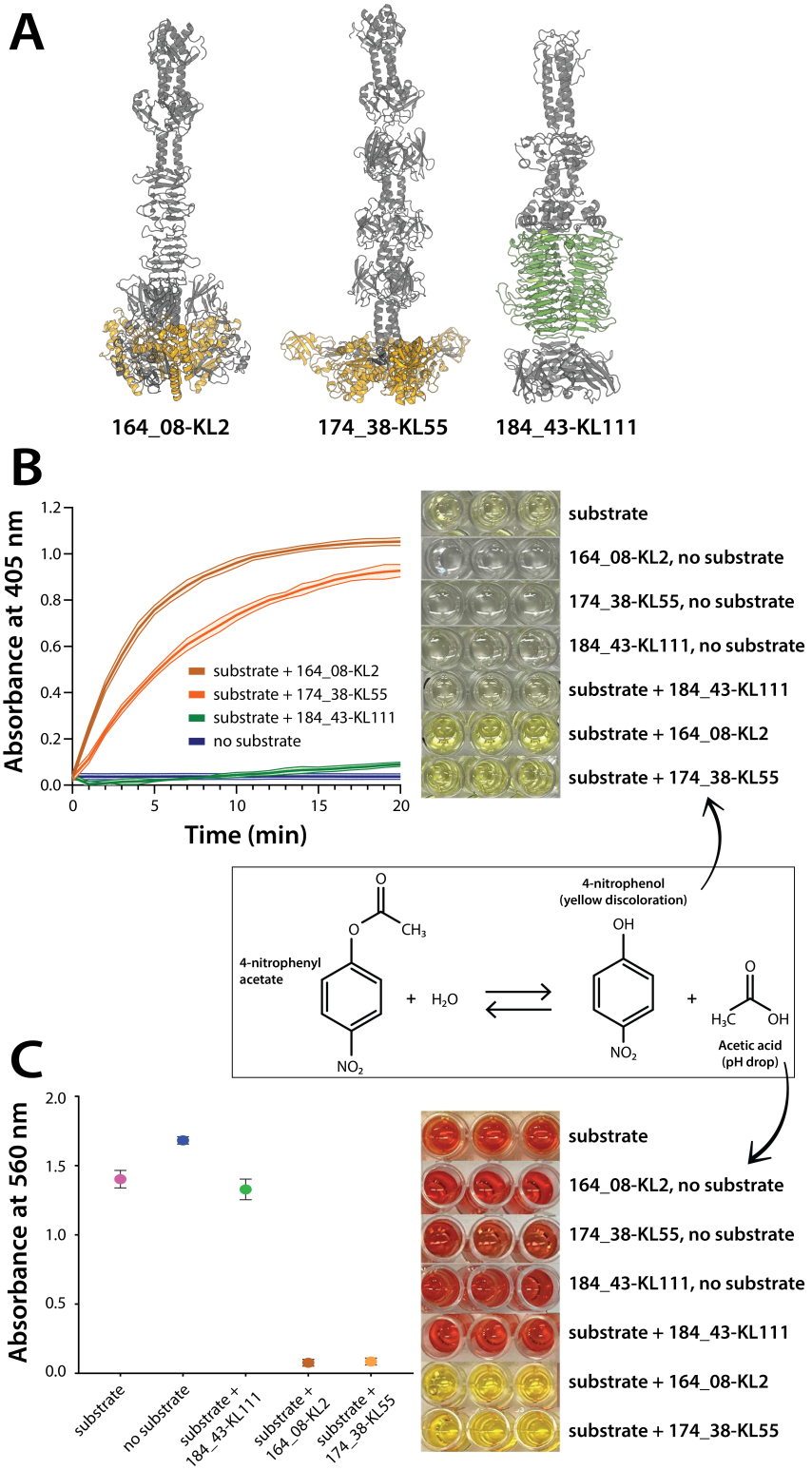
A
164_08-KL2
174_38-KL55
184_43-KL111
B
Absorbance at 405 nm
Time (min)
substrate + 164_08-KL2
substrate + 174_38-KL55
substrate + 184_43-KL111
no substrate
substrate
164_08-KL2, no substrate
174_38-KL55, no substrate
184_43-KL111, no substrate
substrate + 184_43-KL111
substrate + 164_08-KL2
substrate + 174_38-KL55
4-nitrophenyl acetate
4-nitrophenol (yellow discoloration)
Acetic acid (pH drop)
C
Absorbance at 560 nm
substrate
no substrate
substrate + 184_43-KL111
substrate + 164_08-KL2
substrate + 174_38-KL55

Figure 7. Esterase activity of SGNH-containing tail fibres against 4-nitrophenyl acetate. (A) AlphaFold3 models of three recombinantly produced proteins: 164_08-KL2 (GWAS prediction for KL2), 174_38-KL55 (GWAS prediction for KL55) and 184_43-KL111 (GWAS prediction against KL111). Both 164_08-KL2 and 174_38-KL55 are tail fibres with an SGNH hydrolase domain (yellow) and did not show depolymerase activity against any of the 119 reference K-types; 184_43-KL111 was a depolymerase with an expected parallel β-helix domain (green) and was active against the KL111 strain. (B) Time course of absorbance at 405 nm for detecting the formation of the yellow product 4-nitrophenol in reactions containing substrate alone and substrate with all three tested proteins. The panel includes photographs of the corresponding reaction wells after 20 minutes of reaction, showing yellow coloration for 164_08-KL2 and 174_38-KL55. The schematic representation of the hydrolysis reaction generating p-nitrophenol and acetic acid is shown in a black box. (C) Absorbance at 560 nm measured after the addition of phenol red, comparing substrate alone, no-substrate controls, and reactions containing substrate with the three tested proteins. The panel includes photographs of the wells showing red coloration in controls and 184_43-KL111, and yellow coloration for 164_08-KL2 and 174_38-KL55.

Discussion

In this work, we aimed to uncover molecular determinants of capsular specificity in temperate phages of *K. pneumoniae*. Specifically, we aimed to help fill three major knowledge gap areas in the context of *Klebsiella*-phage interactions, namely (i) which viral genes determine capsule specificity, (ii) why depolymerases are difficult to find in prophages and (iii) whether predicted capsule-specific depolymerases are functionally active. To this end, we carried out a genome-wide association study (GWAS) on an ecologically diverse dataset of 3,900 bacterial genomes, looking for genetic predictors from their 8,105 prophage regions that correlate with the K-type predictions (K-loci); independently, we carried out a manual search by recombinantly producing over 60 putative depolymerases and testing against a broad panel of *Klebsiella* K-types. The computational approach provides important insight into the first two of these questions (capsule tropism vs. depolymerases), while the comparison of the computational and experimental parts provides insight into the third question (functional activity).

A major advantage of our GWAS-based approach over many machine-learning methods is that it does not rely on any predefined protein functions. This function-agnostic design allowed us to explore which prophage genes determine capsule specificity and why identifiable depolymerases are so rare. In around 30% of K-loci, the analysis yielded no significant predictors under any threshold, suggesting that no single protein family perfectly predicts the distribution of any K-locus. Even for the K-loci with the strongest statistical signal, none achieved near-perfect classification ($F_1 \approx 1$), with only four exceeding F_1 of 0.8 and most remaining below 0.5. These results mirror those of a complementary study by Concha-Eloko and colleagues[28], whose larger deep-learning framework correctly matched 48/164 depolymerases ($\sim 29\%$) to the correct K-locus using the top-ranked prediction, underscoring that depolymerase sequence alone is often a weak predictor of capsular specificity when inferred from lysogenic genomic data. Among our strongest predictors across all K-loci, we observed a striking diversity of prophage functions, dominated by tailspikes and tail fibres carrying either the classical β-helix depolymerase domain or the SGNH hydrolase domain, alongside previously unrecognised receptor-binding proteins and even putative acetyltransferases. A systematic manual search further revealed that many prophages lacked identifiable RBPs altogether—possibly reflecting rapid sequence diversification or incomplete assembly of tail-fibre genes in short-read genomes. The case of the KL1 locus illustrates this complexity well: prophages in KL1 strains encoded a mosaic of SGNH-domain RBPs, RBP clusters shared across distinct K-loci, and in one case

no RBP gene at all (see S13 Figure). Collectively, these findings indicate that capsule tropism in temperate *Klebsiella* phages is a complex phenotype, shaped by diverse infection strategies in which classical polysaccharide depolymerisation represents only one, albeit major, mechanism of host recognition.

A large set of recombinantly produced depolymerases revealed that only 14 out of 50 proteins from the manual search showed evidence of capsule specificity against one of 119 reference K-types. Of those 14, five were active on a different K-type than that predicted of their host. The comparison of the computational and experimental approach revealed that, where GWAS and manual datasets overlapped, predictions and experimental results were largely concordant, except for two cases: one lacking the C-terminus and another that was expressed but inactive. Importantly, however, of the remaining 36 recombinant proteins, 34 were not overexpressed. Our attempts to optimise expression conditions for 8 representative of those 34 proteins resulted in an improved expression of only one protein, but it remained inactive. We did not find support for the hypothesis that these proteins originate from cryptic prophages because their genomes were complete with no excess of transposons, and many were recently integrated, suggesting they are likely infectious. We also found co-occurrence of active and non-expressed depolymerases in the same prophage genome in three cases. Comparison of predicted structures suggested that the non-expressed depolymerases are not defective, and we identified two pairs of enzymes with near-identical sequences in which one protein was successfully expressed and enzymatically active, while the other failed to overexpress. Taken together, these results indicate that recombinant expression and enzymatic activity of prophage depolymerases are highly sensitive to both sequence and construct context. Consequently, functional activity cannot be reliably inferred from sequence identity or structural prediction alone.

The relatively high frequency of the 'SGNH hydrolase' domain among GWAS predictions underscores its importance as a host-range determinant in temperate phages of *K. pneumoniae*. This prevalence is also reflected by the high proportion of RBPs with this domain across all prophages in our dataset (cf., Figure 6). The SGNH hydrolase domain, encompassing a diverse family of esterases and lipases, has previously been identified as a structural topology in some RBPs of other phages[57], including those infecting *Acinetobacter baumannii*[58] and *E. coli*[53]. Furthermore, Prokhorov and colleagues demonstrated that gp63.1, a SGNH-containing RBP from an *E. coli* phage G7C, was an esterase functioning as a deacetylase, removing an O3-acetyl group from lipopolysaccharide without degrading the O-chain[59] – a reaction essential for G7C's infectivity. To see if the SGNH hydrolases observed in this study might have a similar function, we produced two proteins, classified as strong predictions by GWAS: 164_08 (KL2

prediction) and 174_38 (KL55 prediction). We then confirmed esterase activity for both proteins using a colorimetric assay with p-nitrophenyl acetate, which detects hydrolysis by the release of p-nitrophenol and acetic acid, consistent with an active SGNH-type esterase fold capable of cleaving acetyl esters. These results suggest that deacetylation may represent a common alternative strategy in temperate *Klebsiella* phages to recognise and "shave" polysaccharides from acetyl groups (without depolymerisation) to initiate infection.

The idea that temperate phages of *K. pneumoniae* may often rely on deacetylation to initiate infection is consistent with the fact that O-acetylation of bacterial polysaccharides is common, with multiple lines of evidence. O-acetylation has been detected in many K-types, either as part of the defined repeating-unit structure (e.g., K5, K21b, K22, K33, K54, K55, K58, K59, K82)[60,61,62,63,64,65,66,67,68] or observed in NMR spectra without precise localisation (e.g., K4, K8, K10, K11, K13, K20, K24, K30, K41, K43, K44, K49, K63, K64, K69)[69,70,71,72,73,74,75,76,77,78,79,80,81,82,83]. Likewise, K-loci in *K. pneumoniae* often feature genes annotated as acetyltransferases[17]. Incidentally, we also detected putative acetyltransferases in two of our strongest GWAS hits—despite filtering out short genes—and between 5–30% of prophages per K-locus in our dataset encode proteins with a predicted acetyltransferase function (see S14 Figure), suggesting that sugar modification may be driven – at least partially – by incoming prophages. As the presence or absence of an acetyl group in otherwise identical or nearly-identical polysaccharide structures can be associated with different serological assignment, as in the case of K22/K37[84,62] or K30/K33[63,76], it is plausible that current serotyping and K-locus genotyping underestimate capsule diversity, as O-acetylation introduces biochemical variation that can affect phage recognition, host tropism and immune interactions beyond what is captured by these schemes.

Another interesting point is that, if the dynamic interplay between acetylation and deacetylation shapes capsule tropism of bacterial viruses, this may reflect both a direct mechanism of host recognition by temperate phages and a competitive strategy to modify host surfaces in ways that interfere with other phages, particularly those that rely on unmodified polysaccharides. Similar mechanisms have been reported across Gram-negative bacteria, where O-acetylation of surface polysaccharides alters phage susceptibility: in *E. coli* and *Salmonella enterica* serovar Typhimurium, it modulates sensitivity to superinfecting phages[85,86]; in *Shigella flexneri*, temperate phage-mediated O-acetylation drives serotype conversion and can confer phage resistance[87,88]; and in *Pseudomonas aeruginosa*, phage D3 introduces O-acetyl modifications to prevent superinfection by related phages[89,90]. While in *K. pneumoniae* some

depolymerases may tolerate acetylated substrates[91], others may not, suggesting that these subtle modifications could significantly reshape the phage infectivity landscape. Overall, these findings highlight polysaccharide acetylation and deacetylation as a potentially underappreciated layer of viral specificity and phage–phage competition (e.g., via superinfection exclusion mechanisms) in *K. pneumoniae*.

A key limitation of our study is that the dataset underlying the GWAS was derived from isolates collected in only two countries–Australia and Italy–which inevitably constrains the total genetic diversity represented. However, this focused sampling also minimises biases that often affect public databases, which are typically enriched for clinical and antibiotic-resistant isolates. Our collection was assembled independently of resistance or virulence traits and thus provides a less epidemiologically biased view of prophage diversity. To assess how representative our dataset is of the broader species diversity, we compared its sequence type (ST) distribution to two global *K. pneumoniae* resources: ARMnet, which comprises over 45,000 publicly available genomes with standardised epidemiological typing from Pathogenwatch[92], and KlebNNSsero[93], a geographically broad set of around 2,000 neonatal bloodstream isolates assembled without selection for resistance or virulence (see S15 Figure). We found that 390 of 542 STs (72%) have been detected in at least two countries, and 347 STs (64%) have been reported on at least two continents, including globally distributed lineages such as ST11, ST15 and ST307 . Notably, 382 STs (70%) were detected in at least one additional country beyond Italy and Australia, and 311 STs (57%) were detected on at least one additional continent beyond Europe and Oceania. Altogether, this suggests that our dataset captures a representative cross-section of the species' global diversity while maintaining statistical tractability and biological consistency for GWAS analyses.

Beyond sampling, additional limitations concern the analytical and experimental scope of the study. Our GWAS was restricted to genomes with confidently assigned K-loci, which supports biological relevance but limits statistical power by excluding rarer K-types and other species from the *Klebsiella* genus where the K-typing does not work as well. Moreover, GWAS results are inherently sensitive to parameter choices—including the definition of protein clusters, K-locus prediction accuracy, and population-structure correction—and future work combining larger datasets with tuned parameters will likely recover additional associations. We also considered whether degraded or cryptic prophages might bias GWAS results, as non-functional prophages could accumulate mutations unrelated to capsule tropism. However, when we repeated the analyses using only high-quality prophages (CheckV completeness $\geqslant$ 99%), the results were strongly correlated with those from the full dataset ($r = 0.61 - 0.84$; S16 Figure), indicating

that degraded prophages did not significantly influence the associations. Another key limitation is that our experimental validation focused on depolymerase activity detectable via halo formation. As a result, we may have missed RBPs with enzymatic activity that do not produce visible halos or operate via alternative interaction mechanisms. Future work could address this using complementary biochemical and structural approaches, such as quantitative polysaccharide degradation assays or mass spectrometry–based analyses, although these methods are inherently lower throughput and require purified capsule substrates. Finally, while our findings highlight unexpected complexity in capsule recognition among temperate *Klebsiella* phages, it remains unclear to what extent these insights generalise beyond this species or to other phage lifestyles such as strictly lytic (virulent) phages.

In summary, our combined GWAS and experimental screen in a near-random sampling of *K. pneumoniae* prophages demonstrates that, although classical right-handed β-helix depolymerases do indeed mediate capsule tropism in temperate phages, they are far from the sole mechanism at play. SGNH hydrolase–driven deacetylation emerges as a likely and functionally dominant strategy for capsule recognition and modifying the capsule, revealing a previously unappreciated layer of molecular complexity. These findings uncover hidden diversity in polysaccharide modifications across *K. pneumoniae* populations and carry direct implications for designing more effective phage therapeutics and capsule-based vaccines. More broadly, our work underscores the need to integrate population-scale genomics, structural modeling, and rigorous functional assays to fully capture the evolutionary innovations that temperate phages employ to expand and refine their host range.

Methods

Computational approach

Data

To build a representative genome dataset of *Klebsiella spp.*, we combined two collections. The first collection (KASPAH) consisted of 391 *Klebsiella pneumoniae* species complex isolates (*K. pneumoniae*, *K. quasipneumoniae*, *K. variicola*), obtained from a diagnostic laboratory serving four major hospitals in Melbourne between April 2013 and March 2014[35,36,33]. These isolates were sourced from various clinical samples including pneumonia ($n = 44$), rectal swabs ($n = 82$), throat swabs ($n = 14$), urinary tract infections (UTI; $n = 193$), wound or tissue ($n = 28$), disseminated infections ($n = 25$), and other sources ($n = 5$). Sequencing was performed using short-read Illumina technology for all isolates, while 99 were additionally sequenced using long-read Oxford Nanopore Technology (ONT) to generate hybrid short/long-read finished assemblies[37]. The second collection (KLEBPAVIA) encompasses 3,510 *Klebsiella spp.* isolates chosen from 6,548 samples from diagnostic laboratories in Pavia, Italy, and its surroundings between June 2017 and November 2018[34]. This dataset includes 3482 *Klebsiella spp.* isolates, with an additional 27 *K. quasipneumoniae* isolates not published in Thorpe et al. 2022, as well as a single outgroup isolate originally classified as *Klebsiella* but later identified as *Superficieibacter maynardsmithii*[94]. The full dataset consists of 3,900 isolates of *Klebsiella spp.* and a single outgroup. Importantly, the combined collection consists of diverse *Klebsiella* isolates from multiple ecological niches, hence providing a diverse overview of both capsule and prophage diversity within. All genomes were annotated with the PATRIC server between November 2022 and March 2023 and with the recipe Bacteria / Archaea and Taxonomy ID 573 (*K. pneumoniae*). The list of bacterial genomes is provided in S2 Table.

Inference of the population structure

Core genes were defined on the whole dataset (including the outgroup) using panaroo v1.3.4[95] on 'strict' mode, as genes present in at least 99% of isolates. SNPs were extracted from an alignment of 1668 core genes with snp-sites v2.5.1[96]. The maximum likelihood phylogenetic tree was inferred from those SNPs using IQ-TREE v2.1.4[97] with the following parameters: `-t PARS -ninit 2` and the number of constant sites obtained from snp-sites `-C`. The best-fitting model of sequence evolution was identified by IQTREE ModelFinder as GTR+F+I+R10. Independently, we ran Kleborate v2.4.1[98] to predict *Klebsiella* species

for all 3,900 isolates, yielding 19 different species: *K. pneumoniae, K. variicola, K. quasipneumoniae subsp. quasipneumoniae, K. quasipneumoniae subsp. similipneumoniae, K. oxytoca, K. michiganensis, K. ornithinolytica, K. aerogenes, K. pasteurii, K. grimontii, K. planticola, K. quasivariicola, K. spallanzanii, K. huaxiensis, K. Ka3, K. terrigena, K. Ko10, K. huaxiensisNEW* and *K. Kterr4*. We then used PopPunk v2.6.3 to identify bacterial genome sequence clusters (SCs) within each species, which were used as a proxy for bacterial lineages. Four of these species (*K. Ka3, K. Ko10, K. huaxiensisNEW, K. Kterr4*) are outgroups of other species (respectively for *K. aerogenes, K. huaxiensis, K. huaxiensis and K. quasiterrigena*) and were introducing noise in the clustering, hence were not considered for this analysis. The remaining 3,876 genomes assigned to 15 species were analysed with the same approach as described by Thorpe and colleagues[34]. Briefly, for each species we fit the mixture model of PopPunk v2.6.3[38]. The number of components (k) was chosen based on the scatter plot of core and accessory distances and iteratively improved. We then ran the model refinement module of PopPunk for each species individually. However, in the case of two species (*K. spallanzanii* and *K. quasivariicola*) the refinement introduced noise in the clustering and hence was abandoned. We compared the resulting clusters to those from Thorpe et. al. with the Adjusted Rand Index (ARI) using python `sklearn.metrics.adjusted_rand_score`, for which a value of 1 indicates perfect agreement between the two clusterings and 0 indicates random labelling. We compared the resulting trees with microreact for species with ARI < 0.8, all of them appearing coherent despite the discrepancies. The resulting PopPunk clusters are henceforth referred to as bacterial sequence clusters (SCs).

K-locus identification

We used Kaptive v2.0[17], implemented in Kleborate v2.4.1[98], to predict capsule polysaccharide synthesis loci (K-loci) for all isolates, which are predictive of capsular types (K-types). Species and K-loci were predicted with Kleborate for all isolates. To identify *Klebsiella* species for which K-locus prediction is reliable, we searched for the species in which 80% of isolates or more had a confident K-locus prediction (i.e., assigned 'Good', 'Very Good' or 'Perfect' by Kaptive) and which had more than 5 members. The resulting dataset consisted of 2,527 genomes of isolates belonging to four species from the *K. pneumoniae* species complex (*K. pneumoniae, K. variicola subsp. variicola, K. quasipneumoniae subsp. quasipneumoniae, K. quasipneumoniae subsp. similipneumoniae*), henceforth referred to as the GWAS dataset.

Identification of prophages

To detect prophages in bacterial genomes, we used two complementary tools: VirSorter v1.0.5[39] and PhiSpy v4.2.6[40]. VirSorter predicts clusters of viral genes based on sequence similarity to its database (`-db 1`), though it frequently over-predicts the length of prophage regions. PhiSpy, which combines heuristic methods, including protein length, transcription strand directionality or AT/GC skew, tends to predict shorter prophage regions and may capture other mobile genetic elements. As the goal was to minimise the false-negative rate of prophage prediction to consider as many likely viral proteins in the GWAS approach, we took the union of the prophage regions predicted by both tools, creating a primary detection. Manual investigation of the resulting putative viral regions indicated an occasional presence of poorly annotated receptor-binding proteins. Hence, we extended each primary detection by 2 kb upstream and downstream, forming an extended detection. Following this, we applied CheckV v1.0.1[41] to further refine our prophage predictions. CheckV not only splits merged prophages within extended regions but also trims over-predicted regions while estimating prophage completeness and confidence. To account for a potential loss of receptor-binding proteins during this trimming, we again extended the CheckV-defined prophage boundaries by 2 kb upstream and downstream. We retained only prophages with a completeness of at least 50%, confidence rated as low, medium or high, and a minimum length of 2kb. This resulted in a dataset of 8,103 prophages. The mean number of detected prophages per genome was 3.2, with a maximum of 14 prophages found in a single genome. Of those, 4,598 prophages were of high-completeness ($\geqslant$ 90%), which yields 1.8 high-quality prophages per isolate. We did not detect any prophages in 49 bacterial genomes. The prophage detection workflow is implemented in Snakemake v8.10.0 and is accessible on GitHub at `https://github.com/bioinf-mcb/mgg_prophages`.

Functional annotation

To functionally annotate predicted prophages in bacterial genomes, we predicted Open Reading Frames (ORFs) prodigal-gv (v2.11.0)[99] which were clustered MMseqs2 v13.45111[100] with sensitivity `-s 7.5`, using criteria of at least 80% coverage between query and subject sequences and a minimum of 50% amino acid identity, forming clusters (PC80) to streamline downstream analysis. Next, we aligned these protein clusters against the PHROGs database[42] with HHblits[101] using parameters: `-n 2 -cov 0 -p 0.95`. These alignments were then converted into HMM profiles and searched against the HMM ECOD70 database (version F70_20200207) using HHsearch, retaining hits with a probability of 70% or higher.

Within each PC80, we identified all unique ECOD X-levels, selecting the highest-scoring hit within each level, and reported up to five distinct ECOD T-levels per PC80 cluster based on the highest probability. Finally, we mapped predicted structural topologies onto the proteins to provide a comprehensive functional annotation. Prophage functional annotation is accessible as Snakemake v 8.10.0 workflow on GitHub repository under: `https://github.com/bioinf-mcb/mgg_annotation`.

Genome-wide association studies

To identify statistical associations between bacterial polysaccharide capsules (determined by the predicted K-locus) and protein clusters found in prophages, we conducted a genome-wide association study (GWAS) independently of the functional annotation. Protein clusters were generated by selecting proteins with a minimum length of 300 amino acids and clustering them using MMseqs. We used six different clustering levels, using combinations of identity thresholds (none, 50%, 80%) and coverage thresholds (50%, 80%), with the 50/50 parameter combination by default (PC50 clusters). Rare variants were filtered out by retaining only protein clusters found in at least three bacterial isolates from two different sequence clusters (SCs), resulting in 1,690 PC50 clusters out of an initial set of 3,745 clusters.

GWAS was run separately for each of the 35 K-loci which were found in at least 10 SCs. For each K-locus, we ran pyseer v1.3.11[102], which is an implementation of the elastic-net approach[103], which combines a ridge regression and lasso regression, with a parameter alpha which determines the amount of mixing between the two penalties, such that $\alpha = 0$ results in ridge regression and $\alpha = 1$ results in lasso regression. We used SCs as covariates in the model to account for the population structure, under the assumption that capsule tropism in prophages is primarily shaped by recent horizontal acquisitions rather than deep shared ancestry; this is supported by evidence that most prophages in bacterial genomes are relatively recent acquisitions[104,27].

Given the biological nature of the polysaccharide-RBP specificity problem, we a priori expected that a given K-type would be recognised by a single enzyme family. In this situation, we would expect the associations to be determined by a small number of predictors (ideally a single protein cluster), a situation better suited for lasso-regression which eliminates most weak predictors. However, we also allowed the possibility of multiple predictors due to the potential complexities of the analysis, for example high diversity and mosaicism of RBPs resulting in multiple sequence clusters weakly correlating with the phenotype – a situation better suited for ridge-regression which allows the presence of multiple statistical

predictors. Hence, we ran pyseer in two modes, with the first being lasso-dominated ($\alpha = 0.8$) and the second being ridge-dominated ($\alpha = 0.069$). The inputs to pyseer included: (1) a binary matrix for KL types, indicating the presence or absence of the K-locus in question in all 2,527 bacterial isolates; (2) a binary matrix for the presence or absence of each of the 1,690 prophage PC50 clusters in the same isolates; and (3) a mapping of each bacterial isolate to its SC.

The model outputs a regression coefficient (β) and p-value for each protein cluster, quantifying its statistical association with the K-locus. P-values were Bonferroni-corrected (p-value of $0.05/N$, where $N = 1690$ is the number of PC50 protein clusters). While a high β value indicates strong association, visualisation of these associations across the bacterial phylogeny using Phandango v1.3.1 [49] revealed that strong statistical associations do not necessarily translate to accurate predictions, especially in imbalanced datasets where certain K-loci are relatively rare. We therefore considered all protein clusters with $\beta > 0$ and p-value below the significant threshold as predictors, and evaluated their classification performance for each K-locus separately on the full dataset using precision, recall, F1 score, and Matthews correlation coefficient (MCC):

- **Precision** measures specificity:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

- **Recall** measures sensitivity:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

where:

- **True Positive (TP)**: Protein variant present in an isolate with the corresponding K-locus.
- **False Positive (FP)**: Protein variant present in an isolate without the corresponding K-locus.
- **False Negative (FN)**: Protein variant absent from an isolate with the corresponding K-locus.
- **True Negative (TN)**: Protein variant absent from an isolate without the corresponding K-locus.

Additional metrics included:

- **F1 score** balances precision and recall:

$$\text{F1} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

- **Matthews Correlation Coefficient (MCC)** provides a comprehensive measure of prediction quality, incorporating all four confusion matrix outcomes (TP, FP, TN, FN):

$$\mathrm{MCC} = \frac{\mathrm{TP} \times \mathrm{TN} - \mathrm{FP} \times \mathrm{FN}}{\sqrt{(\mathrm{TP}+\mathrm{FP})(\mathrm{TP}+\mathrm{FN})(\mathrm{TN}+\mathrm{FP})(\mathrm{TN}+\mathrm{FN})}}$$

Precision was particularly critical because we aimed to identify highly specific enzymes targeting distinct polysaccharides. A high precision maximises reliability of predictions, even if this occasionally reduces recall. To assess statistical uncertainty of precision and recall, we performed $n = 100$ bootstrap replicates for each K-locus, resampling bacterial isolates with replacement to recalculate TP, FP, and FN values. We derived 95% confidence intervals from the distribution quantiles of these bootstrapped estimates.

Finally, for each PC50 hit selected from the GWAS analysis, we predicted the functions of its sequence members by mapping them to their corresponding PC80 clusters, where functional annotations had previously been assigned. Since different sequence members within the same PC50 cluster could be associated with different PC80 clusters, each member might carry distinct ECOD T-level annotations (or combinations thereof, given the multi-domain nature of phage RBPs). To provide a single functional prediction for each PC50 cluster, we selected the most frequent combination of ECOD T-levels among its sequence members. Additionally, we independently checked whether any sequence member in the PC50 cluster was annotated with the *Pectin lyase-like fold* (ECOD T-level: 207.2.1) in any PC80 cluster. If this annotation was present, we labelled the PC50 cluster as a putative pectin lyase fold.

Experimental approach

Depolymerase prediction

All prophages detected in the KASPAH-REF dataset were considered if they were classified as high or medium completeness (CheckV criteria; see above). Predicted prophage proteins were clustered using MMseqs2 with parameters set to 90% bidirectional sequence coverage and 50% sequence identity. Functional annotations of protein clusters were predicted by searching representative cluster sequences against the PHROGs and UniClust30 databases using HHblits with default parameters (UniRef30 database, E-value cutoff = 10^{-3}, minimum sequence coverage = 20%). A protein cluster was classified as a putative depolymerase candidate if it met the following criteria: (1) the average length of proteins in the cluster was at least 500 amino acids; (2) the cluster was functionally annotated as either 'tail fiber/fibre' or 'tail

spike'; (3) proteins contained predicted enzymatic domains characteristic of polysaccharide-degrading enzymes—specifically lyases (e.g., hyaluronidases, pectin/pectate lyases, alginate lyases, K5 lyases) or hydrolases (e.g., sialidases, rhamnosialidases, levanases, dextranases, or xylanases).

Domain predictions were carried out by comparing representative sequences against multiple databases, as follows: (a) BLASTp searches against the non-redundant protein sequence (nr) database, using standard parameters (expect threshold = 0.05; word size = 5; matrix = BLOSUM62; gap costs: existence = 11, extension = 1; conditional composition score matrix adjustment); (b) HMMER searches against Reference Proteomes (quick search mode; significance thresholds: sequence E-value = 0.01, hit E-value = 0.03); and (c) HHpred searches against the PDB_mmCIF70 database (HHblits MSA generation from UniRef30; maximum MSA iterations = 3; MSA generation E-value cutoff = 10^{-3}; minimum sequence identity = 0%; minimum coverage = 20%; secondary structure scoring: during alignment; alignment mode: realign with MAC [local:norealign], MAC realignment threshold = 0.3; number of target sequences = 250; minimum probability in hitlist = 20%).

Using a combination of the two approaches, a total of 469 proteins meeting all selection criteria were identified. These proteins were clustered at high coverage using MMseqs2 (80% coverage and 50% identity), obtaining 124 representative protein sequences. These sequences were then structurally modelled using AlphaFold2[50] in the monomer mode. The resulting models were screened for conserved structural features, with particular emphasis on the presence of a parallel β-helix fold characteristic of depolymerases. Structural analysis led to the identification of 50 proteins from 32 clusters, with 22 clusters yielding a single representative and 10 clusters yielding multiple representatives to capture sequence and domain-level variation within those groups (altogether 28). The complete list of these proteins is given in S5A Table, and their genomic coordinates are provided in S6 Table.

Bacterial Strains

K. pneumoniae strains used in this study (S4 Table) originate from collections of the Collection de l'Institut Pasteur (CIP), Paris, France, the National Collection of Type Cultures (NCTC), the UK Health Security Agency (UKHSA), the collection of the Department of Pathogen Biology and Immunology (Wrocław, Poland), and the *Klebsiella* Acquisition Surveillance Project at Alfred Health (KASPAH) from Melbourne, Australia[35,36,33]. Bacteria were grown in Tryptone Soya Broth or Agar (TSB or TSA, Oxoid, Thermo Fisher Scientific, Waltham, MA, USA) at 37°C. The collection includes 134 strains representing

119 distinct serotypes.

The *Escherichia coli* strains (Thermo Fisher Scientific, Waltham, MA, USA) were used for recombinant protein production: One ShotTM TOP10 for cloning and One ShotTM BL21 StarTM (DE3) for recombinant protein expression. They were cultured in LB broth, Miller (Luria-Bertani) (Difco, Becton, Dickinson and Company (BD), Franklin Lakes, NJ, USA) at 37 °C supplemented with 100 µg/mL ampicillin (IBI Scientific, Dubuque, IA, USA).

All bacteria were stored at -70°C in a 20% glycerol-supplemented Trypticase Soy Broth (TSB, Becton Dickinson, and Company, Cockeysville, Md, USA).

Cloning, expression and purification

The genomic DNA from bacterial strains containing prophages encoding putative depolymerase was extracted using a Pure Link Genomic DNA Mini Kit (Thermo Fisher Scientific, Waltham, MA, USA) according to the manufacturer's protocol. The genes were amplified by PCR using the specific primer pairs (Genomed, Warsaw, Poland) and Platinum SuperFi DNA polymerase (Thermo Fisher Scientific, Waltham, MA, USA). The desired products were subsequently cloned into the Champion pET101 expression vector (Thermo Fisher Scientific, Waltham, MA, USA) with a C-terminal His tag (6x) and propagated in *E. coli* One ShotTM TOP10 strain according to the manufacturer's protocol. Plasmid DNA was then purified and the accuracy of the clones was confirmed by sequencing (Genomed, Warsaw, Poland).

Following the transformation, recombinant depolymerases were expressed in *E. coli* One Shot™ BL21 Star™ (DE3) (Thermo Fisher Scientific, USA). Cultures were grown in 500 mL LB broth supplemented with 100 µg/mL ampicillin at 37 °C. In addition to the standard expression conditions (1.0 mM IPTG, 18 °C, $OD_{600} \approx 0.5$, 16 h) – first attempt, an alternative induction strategy was tested (0.5 mM IPTG, 15 °C, $OD_{600} \approx 1.0$, 16 h expression) – second attempt. Cultures were harvested by centrifugation (8000 × g, 15 min, 4 °C), and pellets were resuspended in 15 mL of lysis buffer (500 mM NaCl, 20 mM NaH_2PO_4, pH 7.4). Cells were lysed using three freeze–thaw cycles followed by sonication, and crude lysates were clarified by centrifugation (16,000 × g, 30 min, 4 °C) and filtration through 0.22 µm filters (Sarstedt, Nümbrecht, Germany). Purification of His-tagged proteins was performed using Dynabeads™ His-Tag Isolation and Pulldown (Thermo Fisher Scientific, USA) following the manufacturer's instructions. The quality of the purified proteins was analysed by a sodium dodecyl sulfate–polyacrylamide

gel electrophoresis (SDS-PAGE) using a 12% polyacrylamide gel and Precision Plus Protein All Blue molecular weight standard (Bio-Rad, Hercules, CA, USA).

For all expression conditions, enzymatic activity was evaluated using standard spot assays performed on both crude lysates and purified His-tagged proteins. Crude lysates were tested directly to enable the detection of activity even at very low expression levels. As a negative control, crude lysates from bacteria carrying the empty Champion™ pET101 vector were included in all assays to ensure that halo formation did not originate from host-derived or vector-associated background activity.

In addition to employing standard cloning and expression techniques, the production of selected depolymerases (S5A Table) was outsourced to a commercial service provider (GenScript, China). The genes encoding all the aforementioned proteins were synthesised with codon optimisation for expression in *E. coli*, cloned into pET30a expression vector, and prepared by the company as highly purified protein preparations, each bearing an N-terminal 6×His tag.

Enzymatic functionality assays

Enzymatic activity was assessed using a spot assay on *K. pneumoniae* serotype collections listed in S4 Table. Bacterial cultures grew to an optical density at 600nm ≈ 1.0 and 0.5 mL was spread evenly onto TSA plates. After drying, 10 µL of each recombinant enzyme was spotted on the bacterial lawn. As a negative control, 10 µL of a sample obtained from the expression and purification of an empty vector—processed using the same protocol as for the recombinant proteins was also applied. Following overnight incubation at 37 °C and an additional 24 h incubation at room temperature, plates were examined for the presence of halo zones. The minimum concentration of protein required to produce a detectable halo on the bacterial lawn (minimal halo-forming concentration; MHFC) was established using the method described previously[105].

Esterase activity assay for SGNH-domain proteins

SGNH hydrolases are broadly active on simple acetate esters, and hydrolysis of *p*-nitrophenyl acetate (pNPA) is widely used as a diagnostic assay for SGNH-family esterases, including polysaccharide O-acetylesterases[106,55,54,56]. Hydrolysis of pNPA yields *p*-nitrophenol, which can be monitored spectrophotometrically, and acetic acid, which produces a detectable pH shift using phenol red[107,108].

Protein samples. Two recombinant tail fibres predicted to contain SGNH hydrolase domains (164_08, obtained based on prediction for KL2, and 174_38, obtained based on prediction against KL55) and one depolymerase lacking an SGNH domain (KL111; negative control) were obtained as highly purified preparations from GenScript (Piscataway, NJ, USA). Proteins were stored at –80 °C and diluted in 10 mM Tris–HCl (pH 7.5) immediately before use.

Spectrophotometric pNPA hydrolysis assay. Enzymatic activity was evaluated in 96-well flat-bottom plates using 4 *p*-nitrophenyl acetate (4p-NPA; Sigma-Aldrich) as a chromogenic substrate. Each reaction well contained a final volume of 200 µL composed of 10 mM Tris-HCl buffer (pH 7.5), protein at final amount of 10 µg, and p-NPA at a final concentration of 100 µM. All Reactions were incubated at 25 °C, and the release of *p*-nitrophenol was monitored kinetically at 405 nm for 20 min, with absorbance recorded every minute on a microplate spectrophotometer Varioskan LUX (ThermoFisher Scientific). The colour change to yellow, indicating the release of nitrophenol, was also observed visually after incubation. For controls, wells containing substrate without protein (to correct for spontaneous hydrolysis of p-NPA), substrate with the KL111 depolymerase and KL2 and KL55 SGNH hydrolases without substrate were included.

Phenol-red pH-shift assay. Immediately after completion of the 20-min spectrophotometric measurement, a secondary colorimetric assay was performed to assess acetic acid release during hydrolysis. Phenol red was added to each well to a final concentration of 0.002% (w/v). The pH-dependent colour change of phenol red was then observed visually as a qualitative indicator of hydrolytic activity and measured at 560nm.

Supporting Information

Figure S1. (Top) Distribution of prophage completeness as estimated by CheckV per K-type for the 35 K-types shown in Figure 1. Each point shows the completeness of each prophage detected in an isolate with the corresponding K-type. Box-plots are overlaid to show the median values and quantiles. (Bottom) The same but shown for a single representative per phage variant with the highest value of completeness.

Figure S2. (Left) Full distribution of the number of Sequence Clusters (SCs) per K-type for all K-types in the dataset from Figure 1. (Right) Full distribution of the number of phage variants, obtained at wGRR=0.95, per K-type for all K-types in the dataset.

Figure S3. Relationship between Precision and Recall for the 35 genetic variants from Fig. 2A, i.e., those with F1 $\geqslant$ 0.5 and MCC $\geqslant$ 0.5. The black line shows the best-fit linear model (Pearson correlation) to these data. Colours indicate the corresponding functional predictions obtained using PHROGs. The Pearson correlation coefficient and associated p-value are shown in the plot.

Figure S4. Each row represents one of the 14 recombinant proteins from Figure 4 that showed enzymatic activity against a panel of 119 reference K-types. Columns indicate K-loci corresponding to the bacterial hosts of the prophages from which the proteins originated. Green cells denote K-loci on which the enzyme was active; pink cells indicate the K-locus of the host from which the enzyme's prophage was derived. Rows with only green cells indicate activity against the host's own K-locus, while rows with both pink and green highlight activity on a different K-locus than the prophage's host.

Figure S5. Only a fraction of predicted depolymerases can be overexpressed. Analysis of recombinant depolymerases from connected component 1 in the sequence similarity network (manuscript Figure 4) that did not overexpress. (A) Connected component 1 from the sequence similarity network of recombinant and predicted depolymerases (B) AlphaFold3 homotrimer models of proteins from A which were active or did not overexpress. (C) CheckV and Kaptive metadata for prophages and bacterial K-loci from which the genes encoding putative depolymerase were cloned.

.

Figure S6. Only a fraction of predicted depolymerases can be overexpressed. Analysis of recombinant depolymerases from connected component 4 in the sequence similarity network (manuscript Figure 4) that did not overexpress. (A) Connected component 4 from the sequence similarity network of recombinant and predicted depolymerases. (B) AlphaFold3 homotrimer models of proteins from A which were active or did not overexpress. (C) CheckV and Kaptive metadata for prophages and bacterial K-loci from which the genes encoding putative depolymerase were cloned.

.

Figure S7. Comparison of prophage quality and genomic features between active and inactive depolymerases. We tested whether the lack of detectable activity or expression could be explained by the genomic context of the corresponding prophages. Each depolymerase was classified as expressed & active (dark red) or not expressed or not active (light blue), and the prophage of origin was analysed for several genomic attributes. (A) Prophage completeness estimated by CheckV. (B) Number of *K. pneumoniae* sequence types (STs) containing closely related prophages, used as a proxy for prophage age and vertical inheritance. (C) Size of the prophage cluster (number of similar prophages in the dataset), reflecting the frequency of the element across hosts. (D) Number of annotated transposases per prophage genome, a hallmark of prophage degradation. All tested prophages were complete or near-complete, and no significant differences were observed between active and inactive groups for any feature (Fisher's exact and Wilcoxon tests, $p > 0.1$). These analyses indicate that inactivity or low expression cannot be explained by prophage degradation or domestication.

.

Figure S8. Prophages carrying multiple depolymerases from the manual search. (A) Prophage tail gene modules from which ⩾2 putative depolymerase genes were cloned and tested along with their AlphaFold3 homotrimer models. Colours correspond to active depolymerases (pink), putative depolymerases which did not overexpress (grey), central tail fiber (violet), lysozyme (red), resolvase (orange). (B) CheckV and Kaptive metadata for prophages and bacterial K-loci loci from which the genes encoding putative depolymerase were cloned.

.

Figure S9. Similarity of active and not produced depolymerases. Two examples of protein pairs selected for recombinant overexpression and enzymatic activity testing (A, B). For each pair, the receptor-binding domains share >97% percentage identity in amino-acid sequence, and the corresponding AlphaFold3 homotrimer models are shown. Amino-acid substitutions between proteins in a pair are mapped onto the models and displayed as yellow/orange balls (on each monomer), showing their distribution across multiple parts of the structures.

.

Figure S10. Comparison of the K-loci KL3 and KL146 using Clinker. Each arrow denotes a coding gene, shaded areas show similarity between protein sequences of genes in both loci, and opacity denotes identity weighted by gene overlap. K-locus sequences were obtained from Kaptive.

.

Figure S11. Upset plot showing the number of high-quality (at least 99% complete) prophages encoding each ECOD domain or combination of ECOD domains shown in Fig 5A within PHROG-defined RBPs, stratified by K-locus. The most frequent multi-domain combinations are gp11/gp12 + SGNH hydrolase and pectin-lyase-like + Orf210 N-terminal or gp11/gp12 domains.

Figure S12. Number of high-quality (at least 99% complete) prophages whose RBPs lack any of the domains shown in Fig 5A but encode each ECOD domain up to 2 protein-coding genes upstream or downstream from the protein annotated as RBP.

Figure S13. Tail regions of representative 16 prophages found in *K. pneumoniae* isolates with the KL1 locus. These prophages were chosen from all high-quality (100% completeness and 'high' confidence) prophages in KL1 isolates by accounting for bacterial lineage (SC) and phage variant (wGRR = 0.5 threshold). For clarity, prophage genomes were manually curated to highlight tail gene clusters by removing genes located downstream of the tail length tape measure protein. For two prophages (KVV_B1_PHAGE002_M and KQS_B1_PHAGE024_M) genes upstream of the tail tube protein were also removed. Proteins were annotated by the ECOD with a e-value of 10^{-3}. Tail regions were compared using Clinker, and the protein identity is indicated as in S5 Figure.

Figure S14. Number of high-quality (at least 99% complete) prophages that encode acetyltransferases as detected by PHROG (hit to 'acetyltransferase' or 'O-acetyltransferase') with prob $\geqslant 0.9$ and qcov/scov $\geqslant 0.5$, or by ECOD (hit to any ECOD containing a string `acetyltransf` with prob $\geqslant 0.9$ and scov $\geqslant 0.1$. Yellow shows hits to ECOD-only, orange shows hits to PHROG only, red shows hits to both and grey shows this to neither database.

Figure S15. Global distribution of *K. pneumoniae* sequence types (STs) across the GWAS, KlebNNSsero and Pathogenwatch datasets. (A) Overlap in *K. pneumoniae* STs between the three datasets, showing shared and dataset-unique ST subsets. (B) Jitter plot with boxplots showing, for each of the 405 shared STs, the number of countries in which the ST has been detected (Y-axis) plotted against the number of continents where it occurs (X-axis). (C) Stacked bar plots illustrating the country and continent spread of STs within each subset, normalised by the total number of STs in that subset. (D) Bar plot showing the number of sequence types (STs; $n = 542$) in the GWAS dataset (y-axis) as a function of the number of countries in which they have been reported (x-axis), excluding the two GWAS source countries (Italy and Australia). (E) Analogous bar plot showing the number of STs as a function of the number of continents in which they have been reported, excluding the two GWAS source continents (Europe and Oceania). Across the 542 STs in the GWAS dataset, 390 STs (72%) have been reported in at least two countries and 314 STs (58%) in at least three countries, while 347 STs (64%) have been reported on at least two continents. Notably, 382 STs (70%) were detected in at least one additional country beyond Italy and Australia, and 311 STs (57%) were detected on at least one additional continent beyond Europe and Oceania. Together, these results indicate that the majority of lineages represented in the GWAS dataset are globally distributed.

.

Figure S16. Effect of low-quality prophages on GWAS predictions. Low-quality prophages were defined as CheckV completeness <99% or medium/low CheckV confidence; high-quality prophages had completeness $\geqslant$99% and high confidence. (A) Scatter plots compare, for 35 K-loci, the top F1 scores obtained from GWAS runs on 2 prophage datasets, proteins from high- and low-quality prophages (X-axis) versus proteins from high-quality prophage only (Y-axis), and six protein clustering settings. Each point represents a K-locus; row and column labels correspond to the parameters used for prophage protein clustering – i.e. bidirectional protein sequence coverage (column label 'coverage') and protein amino-acid sequence identity (row label 'identity'); outliers with ΔF1 outside the interquartile range are labelled; dashed lines indicate the diagonal; Pearson correlation coefficients (r) are shown in the top-left corner of each panel. (B) Bar plots show, for each K-locus (X-axis), the top F1 score (Y-axis) across the same six clustering settings, comparing top F1-scores obtained from GWAS run on proteins from high- and low- quality prophages (orange) versus high-quality prophages only (blue). Top GWAS predictors from two prophage datasets show overall positive correlation across all clustering. The exclusion of low-quality prophages for most outliers leads to decrease in F1-score value (eg, KL127).

.

Table S1. List of 58 active depolymerases from previously published studies on virulent phages, based on Cheetham and colleagues[21].

Table S2. List of bacterial genomes used, together with the associated metadata.

Table S3. List of prophage genomes detected in S1, together with the associated metadata.

Table S4. List of 119 *K. pneumoniae* strains from the reference K-type collection used in this study.

Table S5. (A) List of experimentally tested, recombinantly produced proteins, (B) Optimisation of expression and activity for eight selected recombinant proteins, (C) List of expression conditions for *Klebsiella* phage depolymerases used in the literature.

Table S6. Map of proteins from S3 Table onto prophages from S2 Table.

- S1 Text: Methodology and results of the relaxed GWAS approach to select β-helix-containing depolymerase predictions.
- S2 Text: All experimental results.
- S3 Text: Sequence and structural comparison of the GWAS predictions (Text S1) and experimental results (Text S2) based on Figure 4B.

Data Availability

The data underlying this study are publicly available in Figshare. Genome assemblies, input data, and intermediate files required to reproduce the results and figures are available under the following DOI:

- `https://doi.org/10.6084/m9.figshare.29181188`

All scripts, workflows, and code used for data analysis are available in the following GitHub repositories:

- Analysis pipeline: `https://https://github.com/bioinf-mcb/mgg_gwas`
- Figure generation pipeline: `https://github.com/rmostowy/klebs-gwas-figs`

These data and code are sufficient to replicate the findings reported in this manuscript. Upon acceptance of the manuscript, the Figshare datasets will be made publicly accessible. Prior to publication, access for peer review is available via the following private link:

- `https://figshare.com/s/0acd5844d74858c250d6`

Funding

This work was supported by the Polskie Powroty programme of the Narodowa Agencja Wymiany Akademickiej (NAWA) awarded to RM, an Installation Grant from the European Molecular Biology Organisation (EMBO) awarded to RM, and grant no. 2020/38/E/NZ8/00432 from the Narodowe Centrum Nauki (NCN). We gratefully acknowledge the Polish high-performance computing infrastructure PLGrid (HPC centres: CI TASK and ACK Cyfronet AGH) for providing computational resources and support under grants no. PLG/2022/015493, PLG/2022/015858, PLG/2023/016392, PLG/2024/017293, and PLG/2025/018394 awarded to RM. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Author contributions

The following contributions can be distinguished by different authors based on the CRediT (Contributor Roles Taxonomy) system:

- **Conceptualization**: ZDK, RJM
- **Methodology**: JK, BM, ZDK, RJM
- **Software**: JK
- **Validation**: AO, JK, BS, RJM
- **Formal Analysis**: AO, JK, VRP, JL, SO, BS, RJM
- **Investigation**: AO, JK, VRP, EPCR, BS, BM, ZDK, RJM
- **Resources**: KEH, EJF, ZDK, RJM
- **Data Curation**: JK, JL
- **Writing – Original Draft**: BS, ZDK, RJM
- **Writing – Review & Editing**: AO, JK, VRP, JL, SO, KEH, EJF, EPCR, BS, BM, ZDK, RJM
- **Visualization**: AO, JK, BS, RJM
- **Supervision**: BM, ZDK, RJM
- **Project Administration**: BM, ZDK, RJM
- **Funding Acquisition**: RJM

References

[1] Strobel HM, Stuart EC, Meyer JR. A Trait-Based Approach to Predicting Viral Host-Range Evolvability. Annual Review of Virology. 2022 Sep;9(1):139–156.

[2] Tan CCS, van Dorp L, Balloux F. The evolutionary drivers and correlates of viral host jumps. Nature Ecology & Evolution. 2024 May;8(5):960–971.

[3] Meyer JR, Dobias DT, Weitz JS, Barrick JE, Quick RT, Lenski RE. Repeatability and contingency in the evolution of a key innovation in phage lambda. Science (New York, NY). 2012 Jan;335(6067):428–432.

[4] Piel D, Bruto M, Labreuche Y, Blanquart F, Goudenège D, Barcia-Cruz R, et al. Phage-host coevolution in natural populations. Nature Microbiology. 2022 Jul;7(7):1075–1086.

[5] Göller PC, Elsener T, Lorgé D, Radulovic N, Bernardi V, Naumann A, et al. Multi-species host range of staphylococcal phages isolated from wastewater. Nature Communications. 2021 Nov;12(1):6965.

[6] Kauffman KM, Chang WK, Brown JM, Hussain FA, Yang J, Polz MF, et al. Resolving the structure of phage-bacteria interactions in the context of natural diversity. Nature Communications. 2022 Jan;13(1):372.

[7] Strathdee SA, Hatfull GF, Mutalik VK, Schooley RT. Phage therapy: From biological mechanisms to future directions. Cell. 2023 Jan;186(1):17–31.

[8] Skurnik M, Alkalay-Oren S, Boon M, Clokie M, Sicheritz-Pontén T, Dąbrowska K, et al. Phage therapy. Nature Reviews Methods Primers. 2025 Feb;5(1):9. Available from: `https://www.nature.com/articles/s43586-024-00377-5`.

[9] Holtappels D, Alfenas-Zerbini P, Koskella B. Drivers and consequences of bacteriophage host range. FEMS microbiology reviews. 2023 Jul;47(4):fuad038.

[10] de Jonge PA, Nobrega FL, Brouns SJJ, Dutilh BE. Molecular and Evolutionary Determinants of Bacteriophage Host Range. Trends in Microbiology. 2019 Jan;27(1):51–63.

[11] Tesson F, Hervé A, Mordret E, Touchon M, d'Humières C, Cury J, et al. Systematic and quantitative view of the antiviral arsenal of prokaryotes. Nature Communications. 2022 May;13(1):2561.

[12] Lood C, Boeckaerts D, Stock M, De Baets B, Lavigne R, van Noort V, et al. Digital phagograms: predicting phage infectivity through a multilayer machine learning approach. Current Opinion in Virology. 2022 Feb;52:174–181.

[13] Boeckaerts D, Stock M, Ferriol-González C, Oteo-Iglesias J, Sanjuán R, Domingo-Calap P, et al. Prediction of Klebsiella phage-host specificity at the strain level. Nature Communications. 2024 May;15(1):4355.

[14] Gaborieau B, Vaysset H, Tesson F, Charachon I, Dib N, Bernier J, et al. Prediction of strain level phage-host interactions across the Escherichia genus using only genomic information. Nature Microbiology. 2024 Nov;9(11):2847–2861.

[15] Paczosa MK, Mecsas J. Klebsiella pneumoniae: Going on the Offense with a Strong Defense. Microbiology and molecular biology reviews: MMBR. 2016 Sep;80(3):629–661.

[16] Follador R, Heinz E, Wyres KL, Ellington MJ, Kowarik M, Holt KE, et al. The diversity of Klebsiella pneumoniae surface polysaccharides. Microbial Genomics. 2016 Aug;2(8):e000073.

[17] Lam MMC, Wick RR, Judd LM, Holt KE, Wyres KL. Kaptive 2.0: updated capsule and lipopolysaccharide locus typing for the Klebsiella pneumoniae species complex. Microbial Genomics. 2022 Mar;8(3):000800.

[18] Wyres KL, Lam MMC, Holt KE. Population genomics of Klebsiella pneumoniae. Nature Reviews Microbiology. 2020 Jun;18(6):344–359.

[19] Majkowska-Skrobek G, Łątka A, Berisio R, Maciejewska B, Squeglia F, Romano M, et al. Capsule-Targeting Depolymerase, Derived from Klebsiella KP36 Phage, as a Tool for the Development of Anti-Virulent Strategy. Viruses. 2016 Dec;8(12):324.

[20] Latka A, Maciejewska B, Majkowska-Skrobek G, Briers Y, Drulis-Kawa Z. Bacteriophage-encoded virion-associated enzymes to overcome the carbohydrate barriers during the infection process. Applied Microbiology and Biotechnology. 2017 Apr;101(8):3103–3119.

[21] Cheetham MJ, Huo Y, Stroyakovski M, Cheng L, Wan D, Dell A, et al. Specificity and diversity of Klebsiella pneumoniae phage-encoded capsule depolymerases. Essays in Biochemistry. 2024 Dec;68(5):661–677.

[22] Latka A, Leiman PG, Drulis-Kawa Z, Briers Y. Modeling the Architecture of Depolymerase-Containing Receptor Binding Proteins in Klebsiella Phages. Frontiers in Microbiology. 2019;10:2649.

[23] Squeglia F, Maciejewska B, Łątka A, Ruggiero A, Briers Y, Drulis-Kawa Z, et al. Structural and Functional Studies of a Klebsiella Phage Capsule Depolymerase Tailspike: Mechanistic Insights into Capsular Degradation. Structure (London, England: 1993). 2020 Jun;28(6):613–624.e4.

[24] Pan YJ, Lin TL, Chen CC, Tsai YT, Cheng YH, Chen YY, et al. Klebsiella Phage PhiK64-1 Encodes Multiple Depolymerases for Multiple Host Capsular Types. Journal of Virology. 2017 Mar;91(6):e02457–16.

[25] Hsieh PF, Lin HH, Lin TL, Chen YY, Wang JT. Two T7-like Bacteriophages, K5-2 and K5-4, Each Encodes Two Capsule Depolymerases: Isolation and Functional Characterization. Scientific Reports. 2017 Jul;7(1):4624.

[26] de Sousa JAM, Buffet A, Haudiquet M, Rocha EPC, Rendueles O. Modular prophage interactions driven by capsule serotype select for capsule loss under phage predation. The ISME journal. 2020 Dec;14(12):2980–2996.

[27] Haudiquet M, Buffet A, Rendueles O, Rocha EPC. Interplay between the cell envelope and mobile genetic elements shapes gene flow in populations of the nosocomial pathogen Klebsiella pneumoniae. PLoS biology. 2021 Jul;19(7):e3001276.

[28] Concha-Eloko R, Beamud B, Domingo-Calap P, Sanjuán R. Unlocking data in Klebsiella lysogens to predict capsular type-specificity of phage depolymerases. Nature Communications. 2025 Oct;16(1):8798.

[29] Power RA, Parkhill J, de Oliveira T. Microbial genome-wide association studies: lessons from human GWAS. Nature Reviews Genetics. 2017 Jan;18(1):41–50.

[30] Coll F, Phelan J, Hill-Cawthorne GA, Nair MB, Mallard K, Ali S, et al. Genome-wide analysis of multi- and extensively drug-resistant Mycobacterium tuberculosis. Nature Genetics. 2018 Feb;50(2):307–316.

[31] Lees JA, Croucher NJ, Goldblatt D, Nosten F, Parkhill J, Turner C, et al. Genome-wide identification of lineage and locus specific variation associated with pneumococcal carriage duration. eLife. 2017 Jul;6:e26255.

[32] Lees JA, Ferwerda B, Kremer PHC, Wheeler NE, Serón MV, Croucher NJ, et al. Joint sequencing of human and pathogen genomes reveals the genetics of pneumococcal meningitis. Nature Communications. 2019 May;10(1):2176.

[33] Gorrie CL, Mirčeta M, Wick RR, Judd LM, Lam MMC, Gomi R, et al. Genomic dissection of Klebsiella pneumoniae infections in hospital patients reveals insights into an opportunistic pathogen. Nature Communications. 2022 May;13(1):3017.

[34] Thorpe HA, Booton R, Kallonen T, Gibbon MJ, Couto N, Passet V, et al. A large-scale genomic snapshot of Klebsiella spp. isolates in Northern Italy reveals limited transmission between clinical and non-clinical settings. Nature Microbiology. 2022 Dec;7(12):2054–2067.

[35] Gorrie CL, Mirceta M, Wick RR, Edwards DJ, Thomson NR, Strugnell RA, et al. Gastrointestinal Carriage Is a Major Reservoir of Klebsiella pneumoniae Infection in Intensive Care Patients. Clinical Infectious Diseases: An Official Publication of the Infectious Diseases Society of America. 2017 Jul;65(2):208–215.

[36] Gorrie CL, Mirceta M, Wick RR, Judd LM, Wyres KL, Thomson NR, et al. Antimicrobial-Resistant Klebsiella pneumoniae Carriage and Infection in Specialized Geriatric Care Wards Linked to Acquisition in the Referring Hospital. Clinical Infectious Diseases: An Official Publication of the Infectious Diseases Society of America. 2018 Jul;67(2):161–170.

[37] Foster-Nyarko E, Cottingham H, Wick RR, Judd LM, Lam MMC, Wyres KL, et al. Nanopore-only assemblies for genomic surveillance of the global priority drug-resistant pathogen, Klebsiella pneumoniae. Microbial Genomics. 2023 Feb;9(2):mgen000936.

[38] Lees JA, Harris SR, Tonkin-Hill G, Gladstone RA, Lo SW, Weiser JN, et al. Fast and flexible bacterial genomic epidemiology with PopPUNK. Genome Research. 2019 Feb;29(2):304–316.

[39] Roux S, Enault F, Hurwitz BL, Sullivan MB. VirSorter: mining viral signal from microbial genomic data. PeerJ. 2015;3:e985.

[40] Akhter S, Aziz RK, Edwards RA. PhiSpy: a novel algorithm for finding prophages in bacterial genomes that combines similarity- and composition-based strategies. Nucleic Acids Research. 2012 Sep;40(16):e126–e126. Available from: `https://academic.oup.com/nar/article/40/16/e126/1027055`.

[41] Nayfach S, Camargo AP, Schulz F, Eloe-Fadrosh E, Roux S, Kyrpides NC. CheckV assesses the quality and completeness of metagenome-assembled viral genomes. Nature Biotechnology. 2021 May;39(5):578–585.

[42] Terzian P, Olo Ndela E, Galiez C, Lossouarn J, Pérez Bucio RE, Mom R, et al. PHROG: families of prokaryotic virus proteins clustered using remote homology. NAR genomics and bioinformatics. 2021 Sep;3(3):lqab067.

[43] Latka A, Lemire S, Grimon D, Dams D, Maciejewska B, Lu T, et al. Engineering the Modular Receptor-Binding Proteins of Klebsiella Phages Switches Their Capsule Serotype Specificity. mBio. 2021 May;12(3):e00455–21.

[44] Abramson J, Adler J, Dunger J, Evans R, Green T, Pritzel A, et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. Nature. 2024 Jun;630(8016):493–500.

[45] van Kempen M, Kim SS, Tumescheit C, Mirdita M, Lee J, Gilchrist CLM, et al. Fast and accurate protein structure search with Foldseek. Nature Biotechnology. 2024 Feb;42(2):243–246.

[46] Sillitoe I, Bordin N, Dawson N, Waman VP, Ashford P, Scholes HM, et al. CATH: increased structural coverage of functional space. Nucleic Acids Research. 2021 Jan;49(D1):D266–D273.

[47] Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, et al. The Protein Data Bank. Nucleic Acids Research. 2000 Jan;28(1):235–242.

[48] Varadi M, Bertoni D, Magana P, Paramval U, Pidruchna I, Radhakrishnan M, et al. AlphaFold Protein Structure Database in 2024: providing structure coverage for over 214 million protein sequences. Nucleic Acids Research. 2024 Jan;52(D1):D368–D375.

[49] Hadfield J, Croucher NJ, Goater RJ, Abudahab K, Aanensen DM, Harris SR. Phandango: an interactive viewer for bacterial population genomics. Bioinformatics (Oxford, England). 2018 Jan;34(2):292–293.

[50] Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. Nature. 2021 Aug;596(7873):583–589.

[51] Stanton T, Jacobsen M, Gymoese P, Basu S, Chan G, Dangor Z, et al. Technical Report: Accuracy of Klebsiella pneumoniae capsule serotype predictions from whole genome sequences. 2025 Jun;Publisher: Zenodo. Available from: `https://zenodo.org/doi/10.5281/zenodo.15742129`.

[52] Patro LPP, Sudhakar KU, Rathinavelan T. K-PAM: a unified platform to distinguish Klebsiella species K- and O-antigen types, model antigen structures and identify hypervirulent strains. Scientific Reports. 2020 Oct;10(1):16732.

[53] Prokhorov NS, Riccio C, Zdorovenko EL, Shneider MM, Browning C, Knirel YA, et al. Function of bacteriophage G7C esterase tailspike in host cell adsorption. Molecular Microbiology. 2017 Aug;105(3):385–398.

[54] Watanabe M, Fukada H, Inoue H, Ishikawa K. Crystal structure of an acetylesterase from Talaromyces cellulolyticus and the importance of a disulfide bond near the active site. FEBS letters. 2015 May;589(11):1200–1206.

[55] Baker P, Ricer T, Moynihan PJ, Kitova EN, Walvoort MTC, Little DJ, et al. P. aeruginosa SGNH hydrolase-like proteins AlgJ and AlgX have similar topology but separate and distinct roles in alginate acetylation. PLoS pathogens. 2014 Aug;10(8):e1004334.

[56] Anderson AC, Stangherlin S, Pimentel KN, Weadge JT, Clarke AJ. The SGNH hydrolase family: a template for carbohydrate diversity. Glycobiology. 2022 Sep;32(10):826–848.

[57] Dunne M, Prokhorov NS, Loessner MJ, Leiman PG. Reprogramming bacteriophage host range: design principles and strategies for engineering receptor binding proteins. Current Opinion in Biotechnology. 2021 Apr;68:272–281.

[58] Liu S, Lei T, Tan Y, Huang X, Zhao W, Zou H, et al. Discovery, structural characteristics and evolutionary analyses of functional domains in Acinetobacter baumannii phage tail fiber/spike proteins. BMC microbiology. 2025 Feb;25(1):73.

[59] Broeker NK, Barbirz S. Not a barrier but a key: How bacteriophages exploit host's O-antigen as an essential receptor to initiate infection. Molecular Microbiology. 2017 Aug;105(3):353–357.

[60] Van Dam JEG, Van Halbeek H, Kamerling JP, Vliegenthart JFG, Snippe H, Jansze M, et al. A bacteriophage-associated lyase acting on Klebsiella serotype K5 capsular polysaccharide. Carbohydrate Research. 1985 Oct;142(2):338–343. Available from: https://linkinghub.elsevier.com/retrieve/pii/0008621585850370.

[61] Chowdhury TA, Jansson PE, Lindberg B, Lindquist U. Structural studies of an extracellular polysaccharide (K21b) elaborated by Klebsiella type 21b. Carbohydrate Research. 1989 Jul;190(1):145–151. Available from: https://linkinghub.elsevier.com/retrieve/pii/0008621589841540.

[62] Parolis LAS, Parolis H, Niemann H, Stirm S. Primary structure of Klebsiella serotype K22 capsular polysaccharide: Another glycan containing 4-O-[(S-1-carboxyethyl]-d-glucuronic acid. Carbohydrate Research. 1988 Aug;179:301–314. Available from: https://linkinghub.elsevier.com/retrieve/pii/0008621588841260.

[63] Lindberg B, Lindh F, Lönngren J, Sutherland IW. Structural studies of the capsular polysaccharide of Klebsiella type 30. Carbohydrate Research. 1979 Nov;76(1):281–284. Available from: https://linkinghub.elsevier.com/retrieve/pii/0008621579800312.

[64] Dutton GGS, Merrifield EH. The capsular polysaccharide from Klebsiella serotype K54; location of the O-acyl groups, and a revised structure. Carbohydrate Research. 1982;105(2):189–203. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0008621500849678.

[65] Bebault GM, Dutton GGS. Structural investigation of the capsular polysaccharide of Klebsiella serotype 55. Carbohydrate Research. 1978 Jul;64:199–213. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0008621500837015.

[66] Dutton GGS, Savage AV. Structural investigation of the capsular polysaccharide of klebsiella serotype K58. Carbohydrate Research. 1980 Sep;84(2):297–305. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0008621500855597.

[67] Lindberg B, Lönngren J, Rudén U, Nimmich W. Structural studies of the capsular polysaccharide of Klebsiella type 59. Carbohydrate Research. 1975 Jun;42(1):83–93. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0008621500841026.

[68] Jansson PE, Lindberg B, Widmalm G, Dutton GGS, Lim AVS, Sutherland IW. Structural studies of the capsular polysaccharides from Klebsiella types 8 and 82, a reinvestigation. Carbohydrate Research. 1988 Apr;175(1):103–109. Available from: https://linkinghub.elsevier.com/retrieve/pii/0008621588801605.

[69] Merrifield EH, Stephen AM. Structural investigation of Klebsiella K-type 4 capsular polysaccharide. Carbohydrate Research. 1981 Oct;96(1):113–120. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0008621500847011.

[70] Cescutti P, Ravenscroft N, Ng S, Lam Z, Dutton GGS. Structural investigation of the capsular polysaccharide produced by a novel Klebsiella serotype (SK1). Location of O-acetyl substituents using NMR and MS techniques. Carbohydrate Research. 1993 Jun;244(2):325–340. Available from: https://linkinghub.elsevier.com/retrieve/pii/0008621583850113.

[71] Dutton GGS, Ng SK, Parolis LAS, Parolis H, Chakraborty AK. A re-investigation of the structure of the capsular polysaccharide of Klebsiella K10. Carbohydrate Research. 1989 Oct;193:147–155. Available from: https://linkinghub.elsevier.com/retrieve/pii/0008621589851146.

[72] Thurow H, Choy YM, Frank N, Niemann H, Stirm S. The structure of Klebsiella serotype 11 capsular polysaccharide. Carbohydrate Research. 1975 May;41(1):241–255. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0008621500870238.

[73] Niemann H, Frank N, Stirm S. Klebsiella serotype-13 capsular polysaccharide: Primary structure and depolymerization by a bacteriophage-borne glycanase. Carbohydrate Research. 1977 Nov;59(1):165–177. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0008621500833030.

[74] Choy YM, Dutton GGS. Structure of the Capsular Polysaccharide of *Klebsiella* K-type 20. Canadian Journal of Chemistry. 1973 Sep;51(18):3015–3020. Available from: http://www.nrcresearchpress.com/doi/10.1139/v73-449.

[75] Choy YM, Dutton GGS, Zanlungo AM. The Structure of the Capsular Polysaccharide of Klebsiella K-type 24. Canadian Journal of Chemistry. 1973 Jun;51(11):1819–1825. Available from: http://www.nrcresearchpress.com/doi/10.1139/v73-271.

[76] Lindberg B, Lindh F, Lönngren J, Nimmich W. Structural studies of the capsular polysaccharide of Klebsiella type 33. Carbohydrate Research. 1979 Apr;70(1):135–144.

[77] Joseleau JP, Lapeyre M, Vignon M, Dutton GGS. Chemical and n.m.r.-spectroscopic investigation of the capsular polysaccharide of klebsiella serotype k41. Carbohydrate Research. 1978 Nov;67(1):197–212. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0008621500837428.

[78] Edebrink P, Jansson PE, Widmalm G, Nimmich W. The structure of the capsular polysaccharide from Klebsiella K43. Carbohydrate Research. 1994 Apr;257(1):107–115. Available from: https://linkinghub.elsevier.com/retrieve/pii/000862159484111X.

[79] Dutton GGS, Folkman TE. Structural investigation of the capsular polysaccharide of Klebsiella-serotype K44. Carbohydrate Research. 1980 Jan;78(2):305–315. Available from: https://linkinghub.elsevier.com/retrieve/pii/0008621580900117.

[80] Joseleau JP. Structural investigation of the capsular polysaccharide of Klebsiella serotype K 49. Carbohydrate Research. 1985 Oct;142(1):85–92. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0008621500907353.

[81] Joseleau JP, Marais MF. Structure of the capsular polysaccharide of Klebsiella K- type 63. Carbohydrate Research. 1979 Dec;77(1):183–190. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0008621500838045.

[82] Merrifield EH, Stephen AM. Structural studies on the capsular polysaccharide from klebsiella serotype k64. Carbohydrate Research. 1979 Sep;74(1):241–257. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0008621500847801.

[83] Hackland PL, Parolis H, Parolis LAS. A structural investigation of the capsular polysaccharide of Klebsiella K69. Carbohydrate Research. 1988 Feb;172(2):209–216. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0008621500908553.

[84] Lindberg B, Lindqvist B, Lönngren J, Nimmich W. Structural studies of the capsular polysaccharide of Klebsiella type 37. Carbohydrate Research. 1977 Oct;58(2):443–451. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0008621500843712.

[85] Wollin R, Stocker BA, Lindberg AA. Lysogenic conversion of Salmonella typhimurium bacteriophages A3 and A4 consists of O-acetylation of rhamnose of the repeating unit of the O-antigenic polysaccharide chain. Journal of Bacteriology. 1987 Mar;169(3):1003–1009.

[86] Knirel YA, Prokhorov NS, Shashkov AS, Ovchinnikova OG, Zdorovenko EL, Liu B, et al. Variations in O-antigen biosynthesis and O-acetylation associated with altered phage sensitivity in Escherichia coli 4s. Journal of Bacteriology. 2015 Mar;197(5):905–912.

[87] Allison GE, Verma NK. Serotype-converting bacteriophages and O-antigen modification in Shigella flexneri. Trends in Microbiology. 2000 Jan;8(1):17–23.

[88] Teh MY, Furevi A, Widmalm G, Morona R. Influence of Shigella flexneri 2a O Antigen Acetylation on Its Bacteriophage Sf6 Receptor Activity and Bacterial Interaction with Human Cells. Journal of Bacteriology. 2020 Nov;202(24):e00363–20.

[89] Kuzio J, Kropinski AM. O-antigen conversion in Pseudomonas aeruginosa PAO1 by bacteriophage D3. Journal of Bacteriology. 1983 Jul;155(1):203–212.

[90] Kropinski AM. Sequence of the genome of the temperate, serotype-converting, Pseudomonas aeruginosa bacteriophage D3. Journal of Bacteriology. 2000 Nov;182(21):6066–6074.

[91] Ye TJ, Fung KM, Lee IM, Ko TP, Lin CY, Wong CL, et al. Klebsiella pneumoniae K2 capsular polysaccharide degradation by a bacteriophage depolymerase does not require trimer formation. mBio. 2024 Mar;15(3):e0351923.

[92] Argimón S, David S, Underwood A, Abrudan M, Wheeler NE, Kekre M, et al. Rapid Genomic Characterization and Global Surveillance of Klebsiella Using Pathogenwatch. Clinical Infectious Diseases: An Official Publication of the Infectious Diseases Society of America. 2021 Dec;73(Suppl_4):S325–S335.

[93] Stanton TD, Keegan SP, Abdulahi JA, Amulele AV, Bates M, Heinz E, et al.. Distribution of capsule and O types in Klebsiella pneumoniae causing neonatal sepsis in Africa and South Asia: meta-analysis of genome-predicted serotype prevalence to inform potential vaccine coverage. Infectious Diseases (except HIV/AIDS); 2025. Available from: `http://medrxiv.org/lookup/doi/10.1101/2025.06.28.25330253`.

[94] Batisti Biffignandi G, Gibbon MJ, Corbella M, Thorpe HA, Merla C, Castelli M, et al. Genome of Superficieibacter maynardsmithii, a novel, antibiotic susceptible representative of Enterobacteriaceae. G3 (Bethesda, Md). 2021 Feb;11(2):jkab019.

[95] Tonkin-Hill G, MacAlasdair N, Ruis C, Weimann A, Horesh G, Lees JA, et al. Producing polished prokaryotic pangenomes with the Panaroo pipeline. Genome Biology. 2020 Jul;21(1):180.

[96] Page AJ, Taylor B, Delaney AJ, Soares J, Seemann T, Keane JA, et al. SNP-sites: rapid efficient extraction of SNPs from multi-FASTA alignments. Microbial Genomics. 2016 Apr;2(4):e000056.

[97] Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, von Haeseler A, et al. Corrigendum to: IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. Molecular Biology and Evolution. 2020 Aug;37(8):2461.

[98] Lam MMC, Wick RR, Watts SC, Cerdeira LT, Wyres KL, Holt KE. A genomic surveillance framework and genotyping tool for Klebsiella pneumoniae and its related species complex. Nature Communications. 2021 Jul;12(1):4188.

[99] Camargo AP, Roux S, Schulz F, Babinski M, Xu Y, Hu B, et al. Identification of mobile genetic elements with geNomad. Nature Biotechnology. 2024 Aug;42(8):1303–1312.

[100] Mirdita M, Steinegger M, Söding J. MMseqs2 desktop and local web server app for fast, interactive sequence searches. Bioinformatics (Oxford, England). 2019 Aug;35(16):2856–2858.

[101] Steinegger M, Meier M, Mirdita M, Vöhringer H, Haunsberger SJ, Söding J. HH-suite3 for fast remote homology detection and deep protein annotation. BMC bioinformatics. 2019 Sep;20(1):473.

[102] Lees JA, Galardini M, Bentley SD, Weiser JN, Corander J. pyseer: a comprehensive tool for microbial pangenome-wide association studies. Bioinformatics (Oxford, England). 2018 Dec;34(24):4310–4312.

[103] Lees JA, Mai TT, Galardini M, Wheeler NE, Horsfield ST, Parkhill J, et al. Improved Prediction of Bacterial Genotype-Phenotype Associations Using Interpretable Pangenome-Spanning Regressions. mBio. 2020 Jul;11(4):e01344–20.

[104] Bobay LM, Touchon M, Rocha EPC. Pervasive domestication of defective prophages by bacteria. Proceedings of the National Academy of Sciences of the United States of America. 2014 Aug;111(33):12127–12132.

[105] Maciejewska B, Squeglia F, Latka A, Privitera M, Olejniczak S, Switala P, et al. Klebsiella phage KP34gp57 capsular depolymerase structure and function: from a serendipitous finding to the design of active mini-enzymes against K. pneumoniae. mBio. 2023 Oct;14(5):e0132923.

[106] Correia MAS, Prates JAM, Brás J, Fontes CMGA, Newman JA, Lewis RJ, et al. Crystal structure of a cellulosomal family 3 carbohydrate esterase from Clostridium thermocellum provides insights into the mechanism of substrate recognition. Journal of Molecular Biology. 2008 May;379(1):64–72.

[107] Le LTHL, Yoo W, Jeon S, Kim KK, Kim TD. Characterization and Immobilization of a Novel SGNH Family Esterase (LaSGNH1) from Lactobacillus acidophilus NCFM. International Journal of Molecular Sciences. 2019 Dec;21(1):91.

[108] Le LTHL, Yoo W, Jeon S, Lee C, Kim KK, Lee JH, et al. Biodiesel and flavor compound production using a novel promiscuous cold-adapted SGNH-type lipase (HaSGNH1) from the psychrophilic bacterium Halocynthiibacter arcticus. Biotechnology for Biofuels. 2020;13:55.