3 The diversity of type II and type IV toxin-antitoxin systems in the global *K. pneumoniae* population

This chapter is a modified version of the paper "Type II and type IV toxin-antitoxin systems show different evolutionary patterns in the global K. pneumoniae *population" [345]. Cinzia Fino, Matthew Dorman and Alexander Harms conducted the phenotypic experiments for testing the activity of toxins and antitoxins which were selected by me. Leopold Parts, Kenn Gerdes, Eva Heinz and Nicholas Robert Thomson contributed to the research of the original publication. All final language is my own.*

3.1 Introduction

TA systems are bicistronic operons which encode for a toxin, which inhibits cellular processes, and an antitoxin which counteracts the toxins' activity [346], and were introduced in Section 1.3 of this thesis. While TA systems have been well studied in a limited number of laboratory and clinical isolates of *E. coli* [347–350] and *Salmonella enterica sv.* Typhimurium [253], there have been few studies in any bacterium that have considered investigating these systems using large clinically relevant collections. In this chapter, SLING, which was presented in Chapter 2 as a tool to search for operons in large datasets, is used to examine the diversity of TA systems across a collection of *K. pneumoniae* genomes.

Since their first description as plasmid addiction systems, it has become clear that TA systems are ubiquitous across a broad range of prokaryotic plasmids and chromosomes [251,316,335,350–353]. The first study examining the distribution of TA systems on a large scale was conducted in 2005, when Pandey and Gerdes used BLAST to search for TA loci across 126 prokaryotic genomes. It was then revealed that TA systems were highly abundant in the chromosomes of free-living Gram-negative and Gram-positive bacteria [316]. In 2009, Makarova et al. used a guilt-by-association approach to identify novel type II TA combinations across the non-redundant protein and COG databases [251,300]. The distribution of the predicted TAs was examined across large evolutionary scales, and it was revealed that specific TAs were significantly over- or under-represented in various taxa, suggesting different dynamics to their propagation depending on the genetic and ecological backgrounds of their host. Additionally, the distribution of TAs was examined in more detail in a set of 41 closely related prokaryotic genomes. An exceptionally high level of variability in the TA system repertoire was observed, even at these close evolutionary ranges. These results paved the path for future studies, as it was evident that these were highly diverse genetic systems which

have yet to be explored. Since, studies on the distribution of these systems were mostly focused on small high-quality genome collections of reference laboratory strains, which do not necessarily represent the diversity in clinical samples [335,350]. Nonetheless, a study on the distribution of type II TA systems in *E. coli* revealed that these systems were differentially distributed across the *E. coli* phylogroups [350]. A similar study in *K. pneumoniae* revealed that type II TA systems are differentially distributed across *K. pneumoniae* isolates from different sources and across plasmids and chromosomes [335].

The large range of TA systems and their ubiquitous nature across species, plasmids and chromosomes suggest that these elements have an essential role in prokaryotic cell biology, beyond their role in plasmid maintenance. Indeed, they have been implicated in other important cellular processes, many of which contribute to resistance and pathogenicity (See Section 1.3.3). These include the formation of antibiotic-induced persistence [348], defence against bacteriophages, biofilm formation [346,354,355], and through transcriptional read-through, influence the expression of adjoining genes [255]. Therefore, a more systematic approach which examines these systems in a collection of clinically relevant genomes can reveal whether their presence is associated with clinically important genes.

3.2 Aims

The aim of this chapter was to use SLING to systematically analyse the diversity of TA systems in a collection of 259 *K. pneumoniae* isolates. The precise aims of this chapter were:

- Describe the distribution of toxins and their antitoxins in a global and clinically relevant collection of *K. pneumoniae* isolates using SLING.
- Test the activity of predicted toxin and antitoxin pairings
- Examine the connection between the presence of these systems and the presence of clinically important genes including AMR genes and virulence genes.

3.3 Methods

3.3.1 Strains and phylogenetic analysis

Assemblies of 259 *K. pneumoniae* species complex strains taken from [9] were assembled using VELVET (v1.2.07) [356] and annotated using Prokka (v1.5) [293] [357]. The core gene phylogeny was inferred from a core gene alignment generated using Roary [305], and a maximum likelihood tree from the informative SNPs, chosen using SNP-sites [332] (v2.3.2), was constructed using RAxML (v8.2.8) [282] with 100 bootstrap replicates.

3.3.2 Toxin-antitoxin prediction

SLING (v1.1) [311] was used to search for toxins and their cognate antitoxins using the built-in toxin domain database provided in SLING. Please refer to Chapter 2 of this thesis for a complete description of SLING's search strategy. The default structural parameters for a TA search in SLING were applied in the filtering step (minimum toxin length: 30 aa, maximum toxin length: 200 aa, minimum antitoxin length: 50 aa, maximum antitoxin length: 150 aa, maximum overlap between toxin and antitoxin: 20 bp, maximum distance between toxin and antitoxin: 50 bp, order: antitoxin either upstream or downstream to toxin, maximum difference: 100 aa). A cut-off of 75% aa sequence identity was used during the grouping step.

The local sequence identity and alignment coverage per toxin and antitoxin group were extracted from the BLAST+ results from the SLING output. All the antitoxin and toxin sequences from each group were aligned using MUSCLE (v3.8.31) [358]. The global sequence identity was calculated as the pairwise sequence identity between every two sequences in the MSA.

3.3.3 Statistical analysis

Statistical analyses were performed in R (v3.3.1). Toxin and antitoxin accumulation curves were generated using the *specaccum* function in the vegan [359] library with 100 random permutations. PCA was performed using the *prcomp* function. Association between toxins and lineage or the presence of AMR genes, virulence genes or plasmid replicons were performed using Fisher's exact test and corrected for multiple testing using the False Discovery Rate (FDR) with the *p.adjust* function. Differences between groups (*K. pneumoniae* complex species, toxin categories) were assessed using the Wilcoxon test and corrected using FDR. Plotting was done using ggplot2 [360].

3.3.4 Toxin group classification

Toxin groups which were observed in over 80% of isolates of all species were assigned as "ubiquitous". Toxin groups which had at least 4 copies and were found to be significantly associated with *K. pneumoniae* complex species (Fisher's exact test, FDR corrected, $p<0.01$) were assigned "species associated". Toxin groups which were not ubiquitous or species associated were assigned "sporadic" if they had 26 copies or more or otherwise, if they were found to be significantly associated with the presence of AMR genes, virulence genes or plasmid replicons (Fisher's exact test, FDR corrected, $p<0.01$). The remaining toxin groups were assigned "rare".

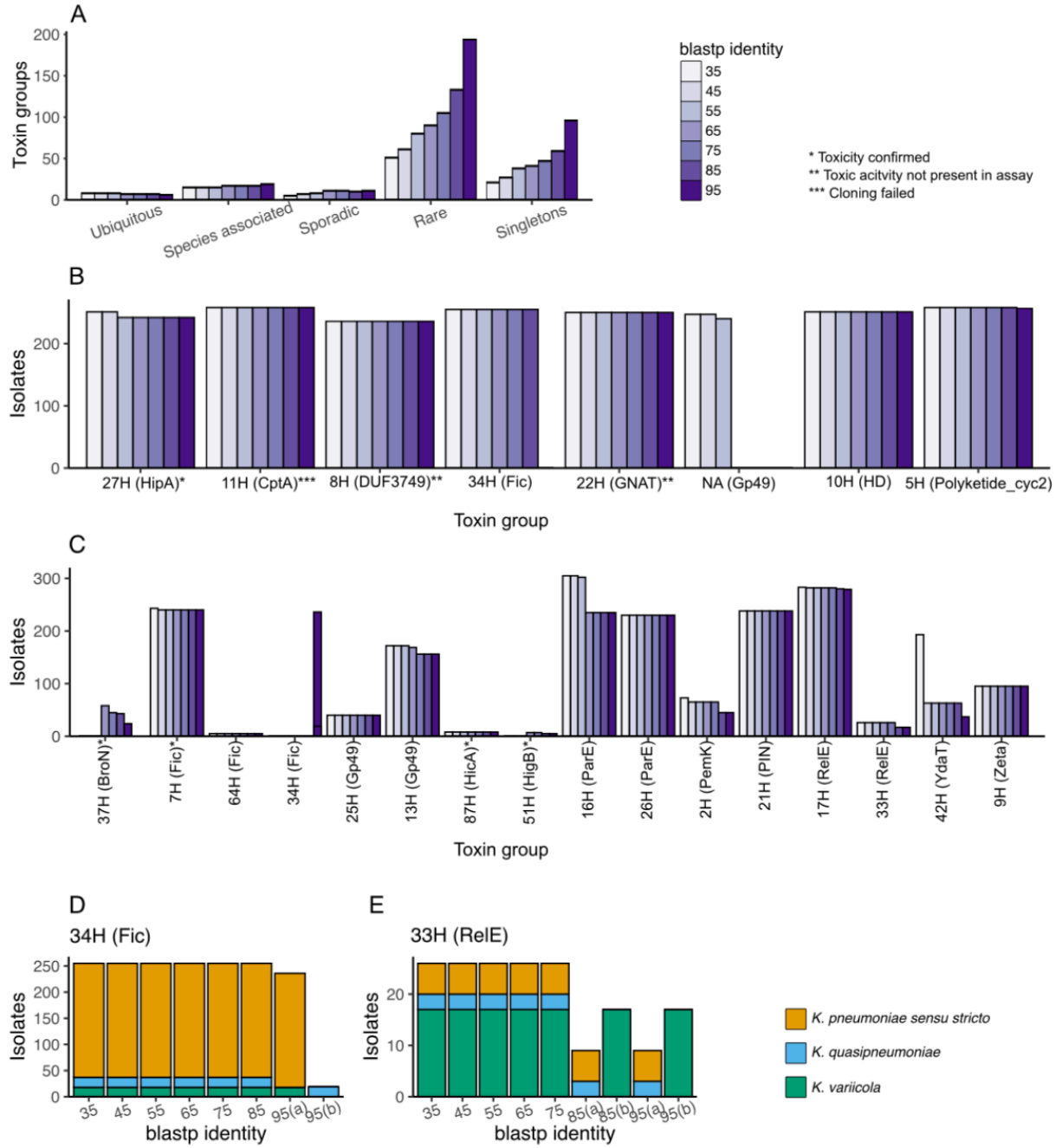

Figure 3.1: Effect of modifying the blastp identity threshold in SLING on the toxin group clustering. A *number of toxin groups from each toxin class for each identity threshold applied. Singletons are toxin groups with only one member.* ***B,D*** *Ubiquitous (**B**) and species associated (**C**) toxin groups under each identity threshold applied. When a bar is missing, the toxin group was not classified as ubiquitous or species associated under the given threshold.* ***D,E*** *Examples of clusterings across thresholds for a ubiquitous toxin group Fic (**D**) and a species associated toxin group RelE_1 (**E**).*

Changing the sequence similarity thresholds for grouping toxins increased the number of toxin groups, however the number of ubiquitous, species-associated and sporadic toxin groups

stayed constant. There was an increase in the number of rare toxin groups which is driven by an increase in the number of singleton toxin sequences (Figure 3.1A). The ubiquitous toxin groups and species-associated toxin groups were robust and stable across all identity thresholds (Figure 3.1B,C). The chosen BLAST identity cut-off of 75% allowed separation of sequences which share similar domains, for instance, DNA binding domains, yet kept homologous sequences together and did not separate sequences by species due to drift (Figure 3.1D,E).

3.3.5 Definition of novel vs known antitoxins

All *in-silico* predicted and experimentally validated type II and IV antitoxin sequences were downloaded from the toxin-antitoxin database TADB (v2, downloaded on 27.08.17) [318,319] and pairwise comparisons between all antitoxin sequences identified by SLING were performed using protein-protein BLAST+ (v2.7) [285]. A SLING antitoxin group was marked as “known” if one or more of the antitoxins in that group shared at least 75% identity and an e-value of 0.01 or lower with an antitoxin from TADB (consistent with the definition of an antitoxin group). Interpro-scan (v5) was used to assign function to the sequences of the novel antitoxins [361]. Sequences which were assigned as antitoxins by Interpro-scan were also marked as “known”. Otherwise, the group was marked as “novel”.

3.3.6 Orphan antitoxins

Antitoxin sequences from an antitoxin cluster were grouped using cd-hit (v4.7) [344] with an identity threshold of 90% and word size of 5 to remove redundant sequences. An antitoxin protein database of the cd-hit representative antitoxins was constructed using BLAST+ (v2.7) [285]. The six frame-translated *K. pneumoniae* genomes from the SLING output [311] were aligned against the antitoxin database using blastn [285]. A CDS was considered an “orphan antitoxin” if a) it was between 50 and 150 aa long, b) it shared 75% sequence identity or more to an antitoxin in the collection and c) the alignment was 50 aa or longer. These settings were chosen to be consistent with the definitions of an antitoxin in the original SLING analysis. The sequences 1,000bp upstream and downstream to the orphan antitoxins were clustered with the respective 1,000bp sequences surrounding the original antitoxin in the viable toxin-antitoxin pair using cd-hit-est with 80% identity threshold and word size of 5. If orphan antitoxin context sequences were in the same cd-hit cluster as the sequences of the original antitoxin, they were marked as “same” and “different” otherwise.

3.3.7 Identification of AMR genes, virulence genes and plasmid replicons

A collection AMR genes were obtained from the modified version of ARG-ANNOT available on the SRST2 website (https://github.com/katholt/srst2/tree/master/data, downloaded on 02.10.16) [288,290]. A dataset of virulence factors was obtained from the *Klebsiella*-specific BIGSDB (http://bigsdb.pasteur.fr/klebsiella/klebsiella.html, downloaded on 22/02/16). The PlasmidFinder database (v1.3) of plasmid replicons was downloaded using ARIBA (v2.12) [283,287]. Presence or absence of a gene in a genome was determined using ARIBA (v2.12) with default settings [283]. Nucleotide-nucleotide BLAST+ (v2.7) of the assemblies against the target gene databases was used to identify contigs which contained a gene of interest (AMR, virulence or plasmid) [285]. A match was determined if any of the associated genes had a BLAST bit score of 200 or more.

3.3.8 Phenotypic testing[5]

Bacterial strains, plasmids, and oligonucleotides used in this study are listed in Appendix A. The sequences of synthesised genes, including mutated ribosomal binding sites and restriction sites where appropriate, are listed in Tables 3.1 and 3.2.

Strains were cultured routinely on lysogeny broth (LB) media. Where appropriate, bacteria harbouring plasmids were cultured on LB media supplemented with 100 µg/ml ampicillin or 30 µg/ml chloramphenicol.

Toxin and antitoxin sequences predicted from computational analysis were synthesised, cloned, and sequence-verified using the GeneArt DNA synthesis service (ThermoFisher Scientific, DE). Toxin sequences were cloned into pNDM220 under P*lac* control [362], and antitoxin sequences into pBAD33 under P*ara* control [363] (Tables 3.1 and 3.2). LB agar plates were supplemented with 1 mM of isopropyl β-D-thiogalactopyranoside (IPTG) for the induction of P*lac* and 0.2% w/v of L-arabinose for the induction of P*araB*. Overnight cultures were washed once and then serially diluted (10^{-1} to 10^{-6}) in sterile phosphate-buffered saline (PBS). 10 µl of the original and diluted cultures (10^{-1} to 10^{-6}) were spotted on LB agar plates containing the induction supplements.

[5] This work was conducted and written by Cinzia Fino, Matthew Dorman and Alexander Harms.

Table 3.1 Phenotypic testing of identified toxins.

Toxin ID	Constuct ID	Pfam domain	Status	Category	TA type	5' Restriction Site	3' Restriction Site
doc	pMJD119	doc	Control - toxic	Control	II	KpnI	KpnI
27H	pMJD127	HipA	Toxic	Ubiquitous	II	KpnI	KpnI
61H	pMJD130	CcdB	Toxic	Sporadic	II	KpnI	KpnI
51H (39P)	pMJD128	HigB	Non-Toxic	Species associated	II	KpnI	KpnI
51H (147P)	pMJD131	HigB	Toxic	Species associated	II	KpnI	KpnI
8H	pMJD121	DUF3749	Non-Toxic	Ubiquitous	II	KpnI	KpnI
87H	pMJD132	HicA	Toxic	Species associated	II	KpnI	KpnI
24H	pMJD134	Gp49	Toxic	Sporadic	II	KpnI	KpnI
72H	pMJD129	HD	Non-Toxic	Sporadic	II	KpnI	KpnI
12H	pMJD122	RES	Non-Toxic	Sporadic	II	KpnI	KpnI
44H	pMJD138/9	ParE	Toxic	Sporadic	II	KpnI	KpnI
14H	pMJD133	Gp49	Toxic	Sporadic	II	KpnI	KpnI
31H	pMJD125	HicA	Toxic	Rare	II	KpnI	KpnI
54H	pNDM_54H	BroN	Non-Toxic	Rare	II	KpnI	KpnI
22H	pMJD124	GNAT	Non-Toxic	Ubiquitous	II	KpnI	KpnI
7H	pMJD120	Fic	Toxic	Species associated	II	KpnI	KpnI
11H	Failed	CptA	Cloning failed	Ubiquitous	IV	KpnI	KpnI
37H	pMJD126	BroN/ANT	Toxic	Species associated	II	KpnI	KpnI
18H	pMJD123	CcdB	Non-Toxic	Sporadic	II	KpnI	KpnI

Table 3.2 Combinations of toxin-antitoxins tested for antitoxin inhibition.

Antitoxin ID	Paired toxins	Operon structure	Status	Novelty	Predicted function	5' Restriction Site	3' Restriction Site
PhD	doc		Control - inhibited	Control	Control	KpnI	HindIII
52P (31H)	31H (HicA)	52P-31H	Inhibited	Novel	Domain of unknown function (DUF1902)	KpnI	HindIII
52P (54H)	31H (HicA)	54H-52P	Inhibited	Novel	Domain of unknown function (DUF1902)	KpnI	HindIII
3P	7H (fic)	3P-7H	Inhibited	Known	Known	KpnI	HindIII
168P	7H (fic)	3P-7H-168P	No inhibition	Novel	Unassigned	KpnI	HindIII
24P	27H (hipA)	24P-27H	Inhibited	Known	Known	KpnI	HindIII
27P	14H (Gp49)	14H-27P	Inhibited	Novel	DNA binding	KpnI	HindIII
23P	44H (ParE)	44P-44H-23P	No inhibition	Novel	Unassigned	KpnI	HindIII
44P	44H (ParE)	44P-44H	Inhibited	Novel	Unassigned	KpnI	HindIII
45P	87H (hicA)	87H-45P	Toxic	Known	Known	KpnI	HindIII
48P	61H (CcdB)	48P-61H	Inhibited	Known	Known	KpnI	HindIII
62P	37H (BroN)	62P-37H	Toxic	Novel	consensus disorder prediction	KpnI	HindIII
26P	37H (BroN)	37H-26P	No inhibition	Novel	Domain of unknown function (DUF4222)	KpnI	HindIII
67P	24H (Gp49)	24H-67P	Partial inhibition	Known	Known	KpnI	HindIII
147P	51H (HigB)	51H-147P	Inhibited	Novel	DNA binding	KpnI	HindIII
39P	51H (HigB)	51H-39P	Inhibited	Novel	DNA binding	KpnI	HindIII

Lyophilised plasmids were rehydrated in nuclease-free water. In order to ensure that *in vitro* validation experiments were performed using a single clone of each synthesised construct, each plasmid was propagated and prepared from a cloning strain of *E. coli*. Briefly, *E. coli* was

cultured aerobically in 100 ml LB broth to an OD_{600} of approximately 0.5 (200 rpm, 37 ˚C). Cells were harvested by centrifugation and resuspended in ice-cold 10 mM calcium chloride ($CaCl_2$) solution. Cells were washed three times in $CaCl_2$ solution, collected by centrifugation, resuspended in 10 mM $CaCl_2$ containing 25% v/v glycerol, and frozen at -80 ˚C. One microlitre of each plasmid solution was used to transform these chemically competent *E. coli* by heat shock (plasmid incubated with bacteria on ice for 30 min, heat shock at 42 ˚C for 30 sec, 5 min immediate recovery on ice). Transformed cells were recovered for one hour at 37 ˚C (200 rpm), and transformants were selected for on solid LB media supplemented with appropriate antibiotics. One colony was picked and single-colony purified; the purified clone was then cultured overnight in 5 ml LB supplemented with antibiotics. Plasmids were extracted from 2 ml of each culture using the QIAprep Spin Miniprep kit (Qiagen, #27104) and the remaining culture was mixed with glycerol (25% v/v final concentration) and stored at -80 ˚C.

3.4 Results

3.4.1 Type II and type IV TA systems are highly abundant in the *K. pneumoniae* species complex

259 *K. pneumoniae* species complex genomes representing the global diversity were included in this study [9] (See Section 3.3.1). These include 222 *K. pneumoniae sensu stricto*, 18 *K. quasipneumoniae* and 19 *K. variicola* isolates (Figure 3.2A), including isolates taken from community and hospital acquired infections, those causing invasive and non-invasive disease and those isolated from both animals and plants [9].

SLING was used to search for TA pairs within our genomic dataset [311]. For clarity, a group of toxins or antitoxins which have been clustered together based on their amino-acid sequence identity are referred to as “toxin group” and “antitoxin group”, respectively. The toxin groups were named by the profile by which they were found.

Using a collection of 55 (52 type II, 3 type IV) Pfam toxin profiles as the input for SLING [311], a total of 140 toxin groups (130 type II, 10 type IV) and 233 antitoxin groups (211 type II, 23 type IV), forming 244 different toxin-antitoxin structures in the genomes included in this study were identified (Appendix B and E). Altogether, TA systems were highly prevalent in all members of the *K. pneumoniae* species complex, with a median of 19 loci per isolate genome (range 11-29, Figure 3.2B). A PCA showed a clear separation into the three species based on toxin repertoire (Figure 3.2C). Furthermore, *K. variicola* had a higher median of 22 TA systems per isolate compared to 18 and 19 in the other two species (Figure 3.2D; pairwise Wilcoxon

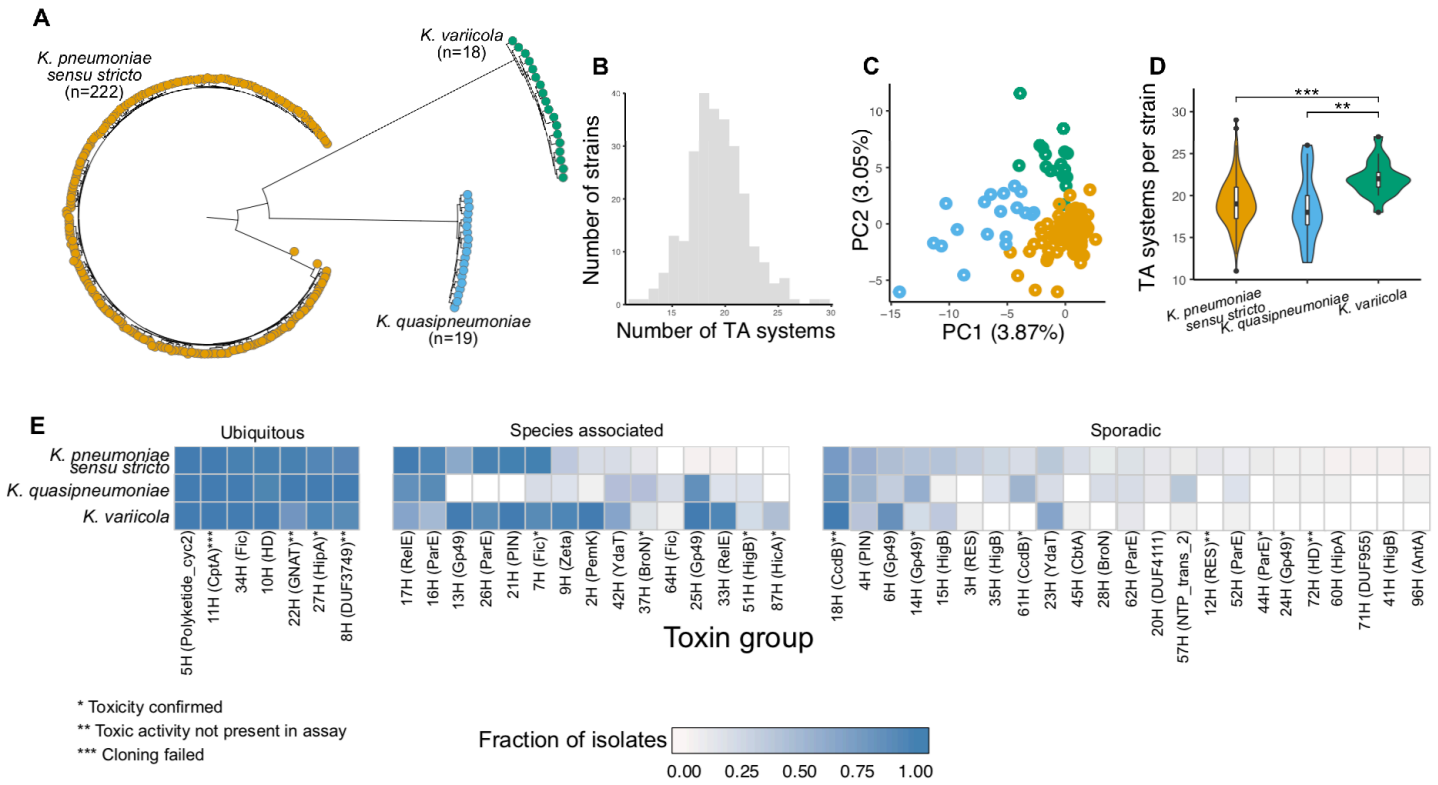

Figure 3.2: Diversity of toxins in* K. pneumoniae *species complex. ***A*** *Core gene phylogeny of the 259 selected* K. pneumoniae *species complex genomes.* ***B*** *Number of predicted TA systems per isolate.* ***C*** *First two principal components of PCA analysis of toxin repertoire coloured by* K. pneumoniae *complex species (yellow:* K. pneumoniae *sensu stricto, blue:* K. quasipneumoniae*, green:* K. variicola*).* ***D*** *Number of predicted TA systems per isolate, stratified by* K. pneumoniae *complex species. Pairwise Wilcoxon rank sum test, *p<0.01, **p<0.001, FDR corrected.* ***E*** *Fraction of isolates from each* K. pneumoniae *complex species possessing each of the toxin groups. Toxin groups are categorised by their distribution patterns (detailed in Appendix B). The toxin Pfam profile used to identify the toxin group is in brackets.*

rank sum test p<0.01, FDR corrected). These figures are slightly higher to those observed in previous studies on TAs in *K. pneumoniae* and *E. coli [335,350]*.

Based on sequence similarity, the number of defined toxin groups per toxin Pfam profile ranged from 1-13 (Figure 3.3). The mean sequence variation within any one toxin group ranged from 68.95-100% local identity at the amino-acid level covering 59.33-100% of the full length of the protein (46.37-100% amino-acid identity over the complete protein) (Appendix B). This highlights the diversity of candidate toxins linked to functionally tested domains that were identified. For instance, the sequences of toxin group 31H containing the HicA domain were aligned to the toxins containing the HicA domain taken from TADB [318,319] (Figure 3.4). While some key residues are conserved throughout, there are considerable variations between the sequences taken from TADB to each other as well as to our predicted toxin.

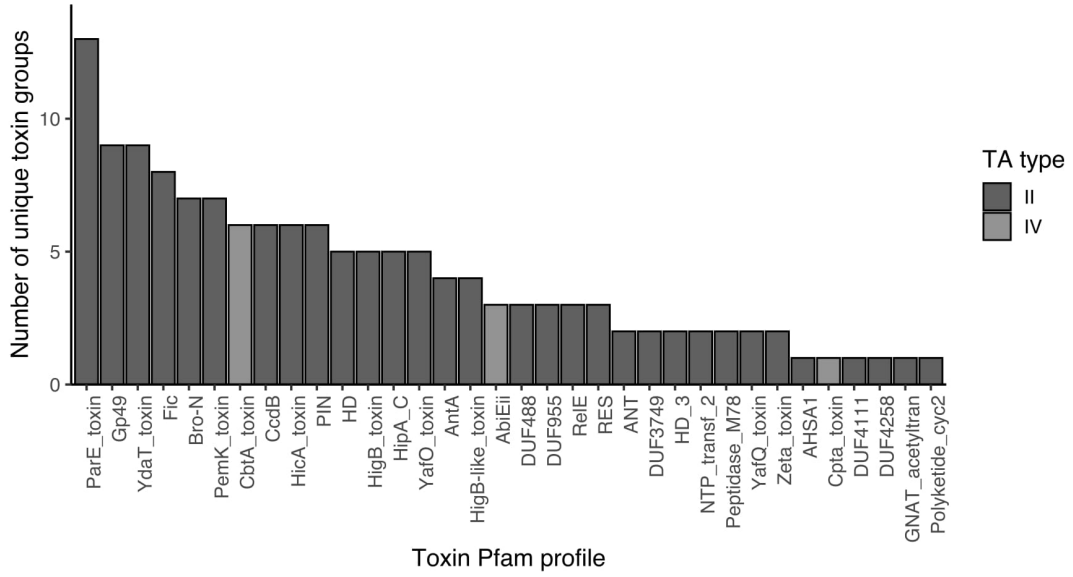

Figure 3.3: Number of unique toxin groups for each of the toxin Pfam profiles used in the search. *Bars are coloured based on the type of TA system the toxin profile is associated with.*

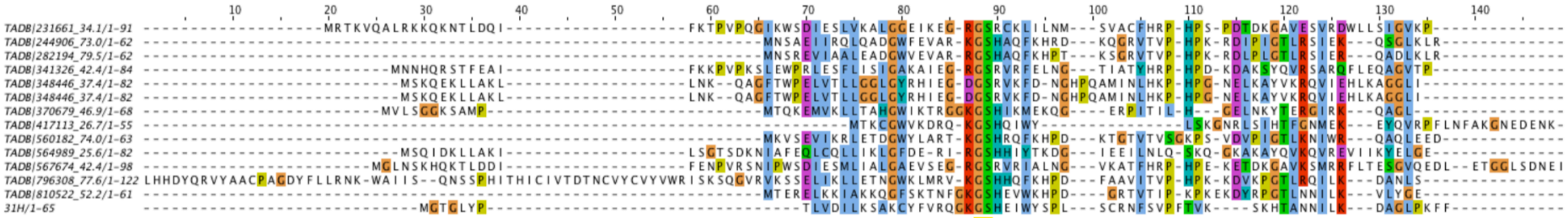

Figure 3.4: Example of diversity of toxins containing a HicA toxin Pfam profile domain. *MSA of all the toxin sequences from TADB containing the HicA toxin Pfam profile, and a representative of toxin group 31H, containing the HicA domain in the* K. pneumoniae *dataset. Alignment was produced using mafft (v7.205) [364]. Image was produced using JalView (v.210) [365].*

3.4.2 Redefining toxins based on their distribution patterns

The 140 identified toxin groups were categorised into four categories based on their distribution patterns in the dataset (See Section 3.3.4) (Figure 3.2E, Appendix B). Seven toxin groups were ubiquitous (one type IV), present in over 80% of the isolates included in this study and from all three species. Fifteen toxin groups, all type II toxins, differed in prevalence between the three species (Fisher's exact test $p<0.01$, FDR corrected, Figure 3.2E). Twenty-three toxin groups (one type IV) (17%) were distributed sporadically with no species association, including a number which were associated with clinically relevant genes. Finally, the remaining 95 toxin groups (eight type IV) (68%) were rare and found in fewer than 10% of the isolates (Appendix B).

Within the ubiquitous toxin groups, we observed significantly higher nucleotide identity for toxins within the same species compared to toxins from other species (median 99.4% compared to 93.51%, Wilcoxon rank sum test, $p<0.001$, Figure 3.5). The median nucleotide identity for sporadic toxin groups for toxins within a species was 97.06% compared to 96.57% between species. This elucidates the evolution of the ubiquitous toxin groups due to genetic drift within a specific member of the species complex, compared to the likely mobile, sporadic toxin groups where this effect was not observed.

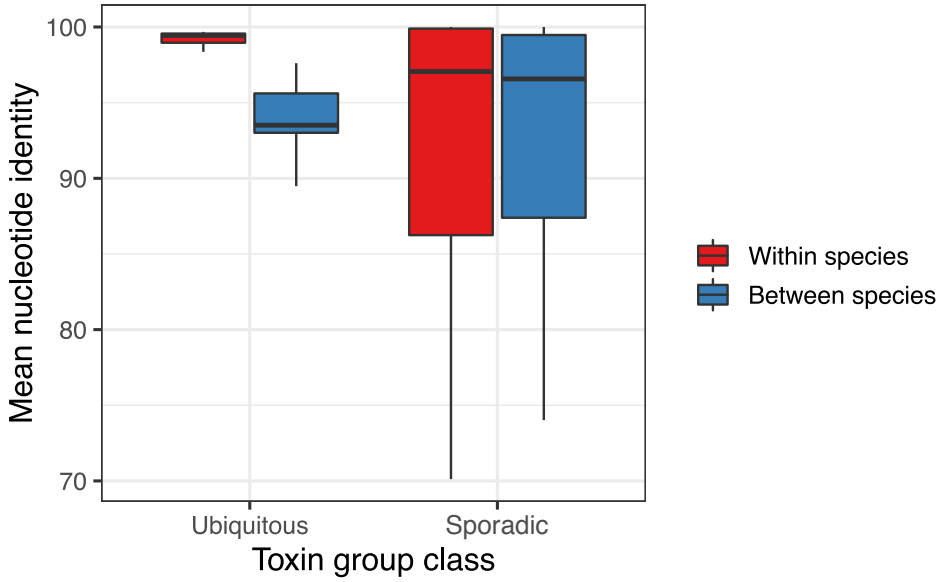

Figure 3.5: Nucleotide identity of toxins within and between species. *Mean nucleotide identity between toxins originating from the same* K. pneumoniae *species and from different* K. pneumoniae *species for all the ubiquitous and sporadic toxin groups.*

The seven ubiquitous toxin groups are known to inhibit translation via mechanisms that do not include RNA cleavage: toxin group 5H (polyketide_cyc) is a homolog of the RatA toxin in *E. coli* which inhibits translation by binding to the 50S ribosomal subunit [338]. Similarly, toxin group 34H (Fic) is a Doc toxin which inhibits translation by phosphorylating and concomitantly inactivating elongation factor TU (EF-Tu) [245]. Toxin groups 22H and 8H with the GNAT and DUF3749 domains are acetyltransferases known to inhibit translation by acetylating aminoacyl-tRNA [366,367]. Group 27H contains a HipA domain which is well described for its association with the high persister phenotype [348,368] and inhibits translation by phosphorylating and concomitantly inactivating glutamyl-tRNA synthetase [369]. Toxin group 11H with the CptA domain belongs to type IV TA system which inhibits cytoskeleton assembly [370]. Finally, group 10H with the HD domain is a phosphohydrolase which is a putative toxin domain from TADB but its exact function is unknown [311,318,319].

The species associated toxin groups presented different distribution patterns across the three *K. pneumoniae* complex species included in this study. *K. pneumoniae sensu stricto* possessed three toxin groups in lower prevalence compared to the other two species (51H (HigB), 64H(Fic) and 25H (Gp49)) (Figure 3.2E). *K. variicola* possessed five toxin groups in higher prevalence compared to *K. pneumoniae sensu stricto* and *K. quasipneumoniae* (42H (YdaT), 9H (Zeta), 2H (PemK), 33H (RelE) and 87H (HicA) domains). Toxin group 87H (HicA) was specific to *K. variicola* and was not observed in the other two species in the dataset. On the other hand, toxin groups 16H (ParE) and 17H (RelE) domains were less common in *K. variicola.* Finally, *K. quasipneumoniae* lacked three toxin groups (21H (PIN), 26H (ParE) and 13H (Gp49)), and rarely possessed toxin group 7H (Fic). On the other hand, toxin group 37H (BroN) was observed in higher prevalence in *K. quasipneumoniae* relative to the other two species. Of these *K. quasipneumoniae* isolates, 11% possessed three copies of this toxin group and 16% possessed two copies (Figure 3.6).

3.4.3 Prediction of novel antitoxins

Accumulation curves of the unique toxin and antitoxin groups identified using SLING suggested that sampling additional *K. pneumoniae* species complex genomes would lead to further identification of new candidate antitoxins (Figure 3.7A). To assess whether the identified antitoxins were known or novel, their sequences were aligned against all type II and type IV antitoxin sequences retrieved from the TADB database [318,319] (See Section 3.3.5). 195 (173 type II, 22 type IV) of the 233 (211 type II, 23 type IV) antitoxins detected in this study were not identified in TADB and were seen to be novel candidate antitoxins linked to a known toxin (Appendix C). For completeness, a predicted function was assigned to the 195 novel

antitoxin groups using interpro-scan (Appendix C) [361]. 19 additional antitoxin groups were matched to known antitoxins by interpro-scan which were not in TADB (antitoxins of toxin profiles YdaT (8), CbtA (4), CcdB (2), Fic (1), PemK (1), PIN (1), HigB (1) and HicA (1)), leading to a final count of 176 novel antitoxins (76%).

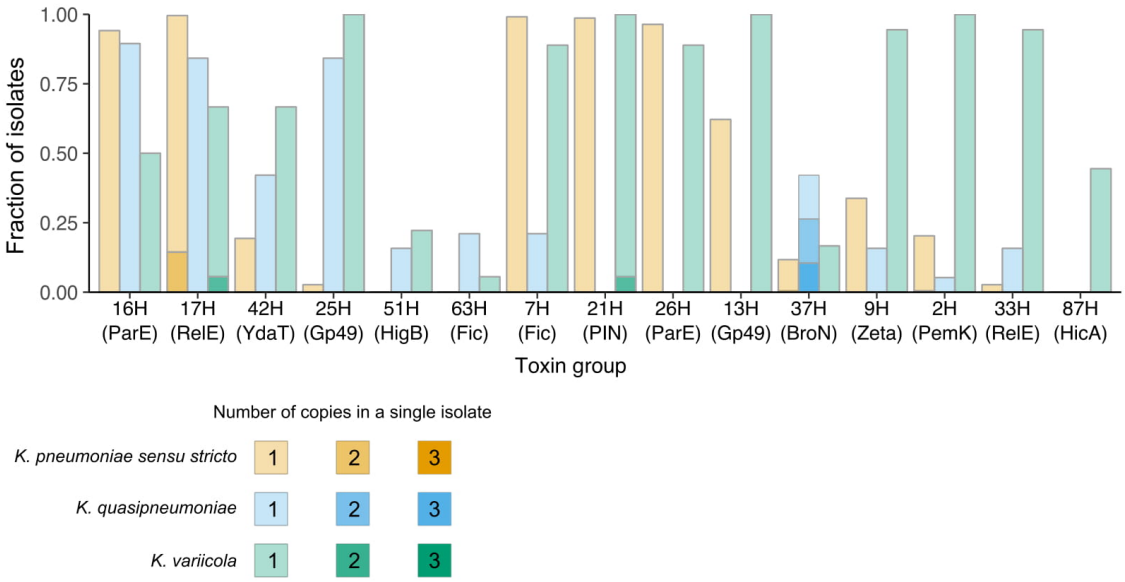

Figure 3.6: Copy number of species-associated toxins. *Fraction of isolates of each of the* K. pneumoniae *complex species possessing each of the species associated toxin groups. Darker shades indicate multiple copies of the toxin group present in an isolate of a species*.

72% of novel antitoxins (127/176) could not be assigned a putative function (Appendix C). Five groups contained one of the toxin profiles used in the toxin search and are the result of disrupted toxins. Twelve groups were predicted to be DNA binding or transcriptional regulators which are plausible functions for antitoxins due to the auto-regulation of the TA operon through conditional cooperativity [346,371]. Another 12 groups were assigned to be intrinsically disordered proteins [372]. The remaining groups contained profiles indicating other functions such as domains of unknown function, ABC transporters, prophages and other functional categories (Appendix C).

For each of the toxin groups, the arrangement of the linked antitoxin was examined: upstream of the toxin (denoted AT-T) or downstream of it (denoted T-AT) (Figure 3.7B). 72% of the known antitoxins were located upstream of the toxin compared to 50% of the novel antitoxins, i.e. novel antitoxin candidates were more commonly located downstream of the toxin relative to the known antitoxins (p = 0.007, Chi squared test).

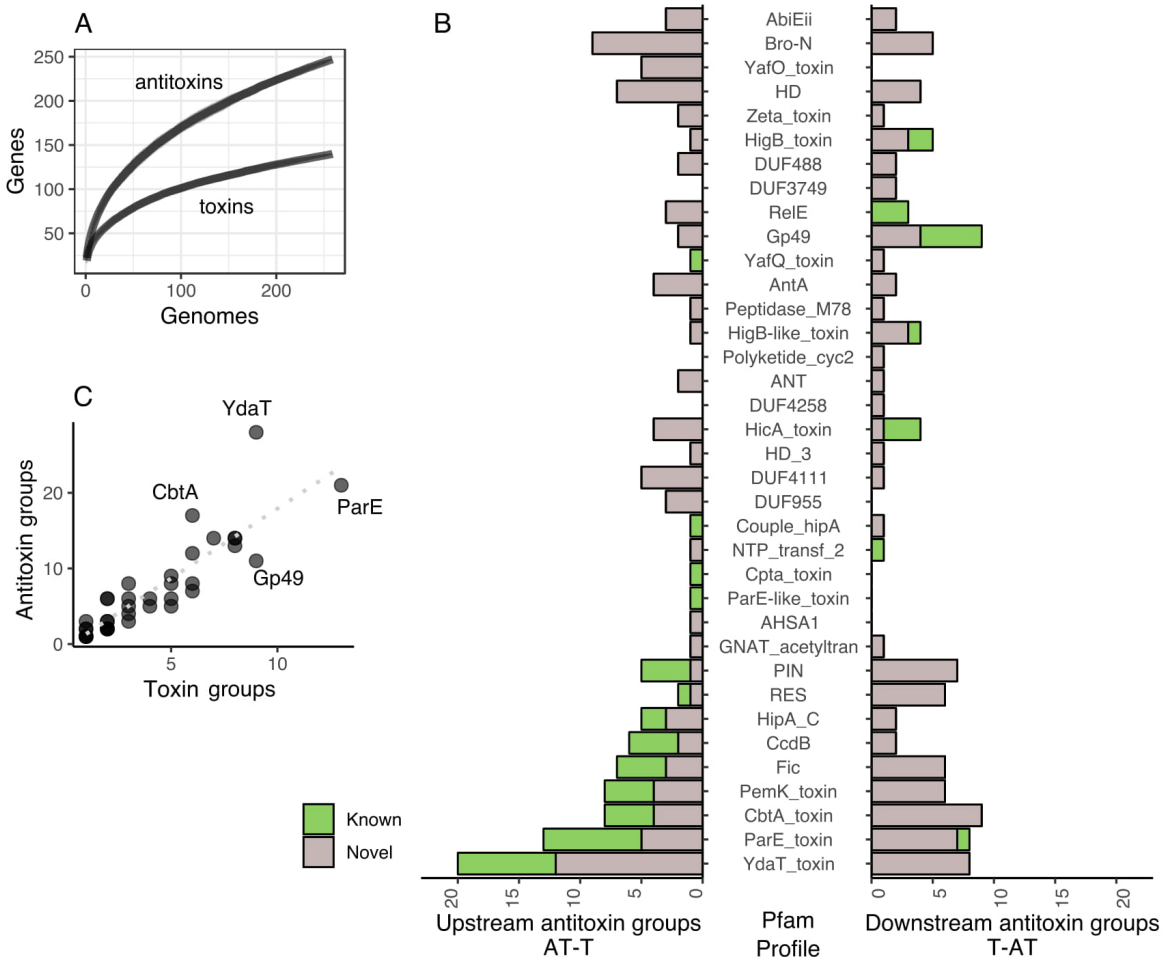

Figure 3.7: Identification of novel antitoxins in the* K. pneumoniae *genomes. A *Accumulation curves of unique toxin and antitoxin groups found in an increasing collection of K. pneumoniae genomes.* ***B*** *Number of antitoxin groups found only upstream (AT-T) and downstream (T-AT) relative to each toxin Pfam profile, coloured by known or novel.* ***C*** *Number of toxin groups of each toxin Pfam profile, relative to the number of antitoxin groups found in their proximity.*

3.4.4 Fluid association and distribution of toxin-antitoxin pairings

Looking at the association between specific toxins and antitoxins we found that with a greater number and diversity of defined toxin groups belonging to the Pfam profile used to search for the toxins, there were concomitantly more antitoxin groups linked to those toxins (0.88 Pearson correlation, 3.7C). The exceptions included the YdaT domain which was found with 28 candidate antitoxin groups and linked to only 9 toxin groups. This both suggests there is coevolution of TA pairs along with instances where a range of different antitoxins can inhibit the same toxin.

We found that a single toxin group can be found with up to a maximum of 12 discrete antitoxins, highlighting the “mix and match” nature of toxin-antitoxin associations [317]. It is important to note that the antitoxin groups are substantially different from each other as a cut off of 75% local amino-acid sequence identity was applied for two antitoxins to be in the same group. Furthermore, the mean sequence variation within any one antitoxin group ranged from 74.64-100% local identity at the amino-acid level covering 61-100% of the alignment length (59.88-100% aa identity over the complete protein), highlighting further the diversity in the candidate antitoxins identified (Appendix C).

In addition to a range of different antitoxins paired to the same toxin, toxins showed a range of operon structures (Figure 3.8A); some toxin groups were linked to a single antitoxin in a conserved position either upstream or downstream of the toxin. Other toxin groups were found in multiple arrangements with the antitoxin sequence and/or location of the antitoxin relative to the toxin changing (Figure 3.8B-H). For the ubiquitous toxin groups, three groups were found in a single arrangement (groups 11H (CptA, a type IV toxin), 5H (polyketide_cyc) and 8H (DUF3749)) (Figure 3.8B). Three other toxin groups (groups 22H (GNAT), 34H(Fic) and 27H(HipA)) were observed in two or three structures often with one structure dominating (>90% of isolates) and the others being rare occurrences of the other structures (<3% of isolates, Figure 3.8C-D). Although the HD toxin group was classified as ubiquitous, one TA arrangement, observed in 80% of isolates, was specific to *K. pneumoniae sensu stricto,* missing in *K. variicola* and replaced by a structure specific to *K. variicola* (Figure 3.8D).

The species-associated toxin group 7H (Fic), was observed in one arrangement which was specific to *K. variicola* (Figure 3.8E). Toxin group 51H (HigB) was associated with two unique antitoxins with one being specific to *K. quasipneumoniae* (Figure 3.8F). Alternatively, other toxin groups possessed multiple operon structures with no clear species association, for instance, toxin group 42H (YdaT) was observed with seven antitoxin groups in eight different arrangements (Figure 3.8G). Other than in a single case (18H (CcdB)), the sporadically distributed toxins were not seen in species-specific arrangements emphasising they are unlikely to be vertically inherited (Figure 3.8H).

Most of the antitoxins identified were toxin group specific. However, antitoxin group 52P was observed with toxin group 31H (HicA) in seven isolates and with toxin group 54H (BroN) in a single isolate. Interestingly, it was always observed upstream to the 31H (HicA) toxin and downstream of 54H (BroN) toxin. The antitoxin proximate to 31H (HicA) shared 83.2% amino acid sequence identity with the antitoxin proximate to 54H (BroN) antitoxin. This antitoxin was

not found in TADB but encodes for a domain of unknown function DUF1902 (PF08972) which is in the same Pfam clan as many other antitoxins (Met_repress, CL0057).

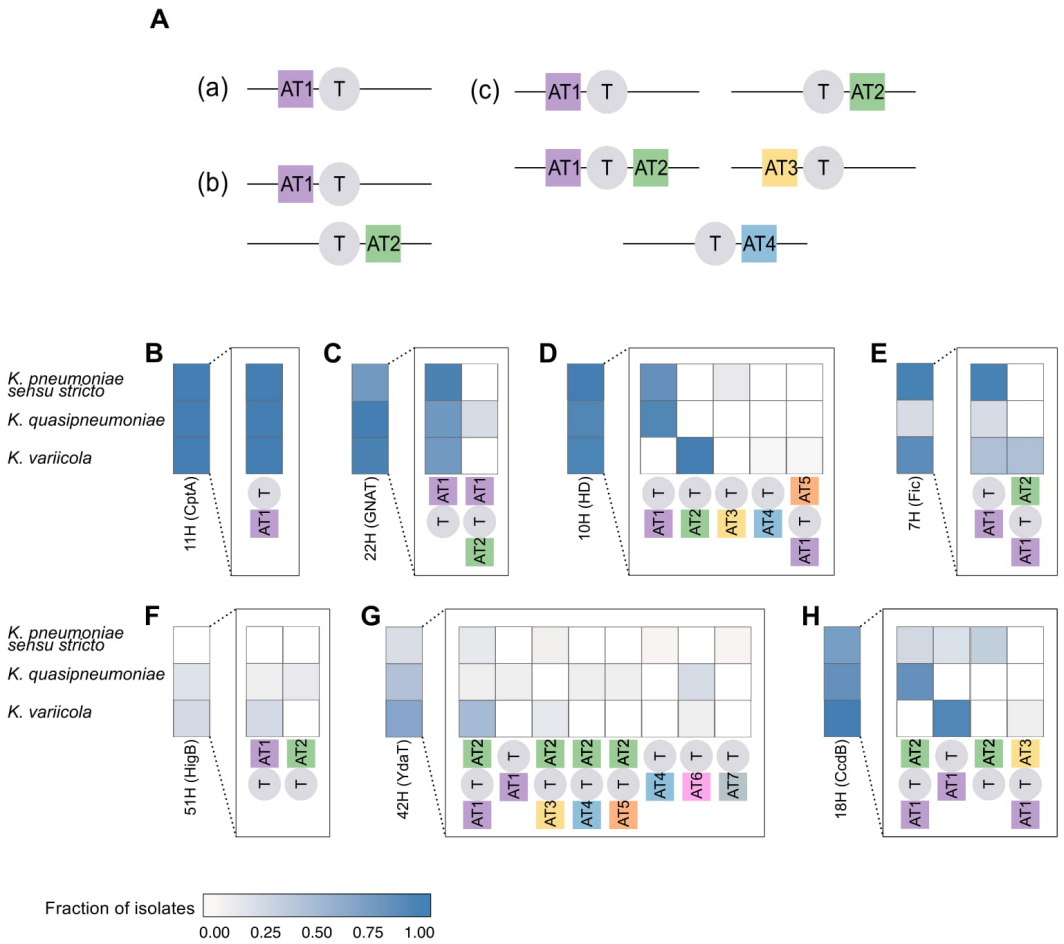

Figure 3.8: Diversity in the observed operon structures for the different toxin categories. A *Examples of range of antitoxins and possible operon structures for a toxin (a) toxin group found in a single structure with a single antitoxin group (b) toxin group found in two different structures with two different antitoxin groups (c) toxin group found in five different structures with four different antitoxin groups.* ***B-H*** *Fraction of isolates from each* K. pneumoniae *complex species possessing each of the operon structures of seven example toxin groups: (**B-D**) ubiquitous, (**E-F**) species associated, (**H**) sporadically distributed.*

3.4.5 Phenotypic testing *in silico* predictions of toxins and confirmation of novel antitoxins[6]

[6] This work was conducted by Cinzia Fino, Matthew Dorman and Alexander Harms.

Due to the apparent diversity of TA systems within and between species and the novel combinations of toxin and antitoxins found in this study, 17 candidate toxins, representing the diversity of toxins within a given group and from a range of genomic backgrounds, were tested for their ability to inhibit bacterial growth in an *Escherichia coli* model system (See Section 3.3.8). Selected were: four ubiquitous, four species associated, seven sporadically distributed and two rare candidate toxins (Table 3.1, Figure 3.9).

The toxicity of all the species associated toxins that were tested was confirmed (groups 51H (HigB), 7H (Fic), 87H (HicA) and 37H (BroN)) (Figure 3.9). Of the remaining toxins, toxicity was observed for the 27H (HipA) toxin group which is ubiquitous across the species complex as well as four of the seven sporadically distributed toxins tested from groups 14H (Gp49), 24H (Gp49), 61H (CcdB), 44H (ParE), and a rare toxin from the 31H (HicA) group. The ubiquitous type IV toxin we tested, 11H ((CptA)), could not be successfully synthesised or cloned, likely due to its toxic activity. The rest of the toxins tested showed no toxic activity under the conditions tested in our assay (summarised in Table 3.1).

Subsequently, 14 candidate antitoxins were tested for their ability to counteract the toxicity of their cognate toxin in the *E. coli* model system (including 10 novel antitoxins; this study; Figure 3.10; Table 3.2). Eight of the fourteen antitoxins (57%) led to complete inhibition of the toxic activity, five of which were novel antitoxins. Three of the confirmed novel antitoxins were predicted to contain DNA binding domains by interpro-scan (39P, 27P, 147P). One antitoxin contained a domain of unknown function (52P) and the final antitoxin did not match any existing entry in Interpro (44P). Three of the confirmed antitoxins in the T-AT format were located downstream of the toxin (groups 27P (Gp49), 147P (HigB) and 39P (HigB)). An additional known antitoxin only partially inhibited toxicity (67P).

For completeness, for operons that had the structure AT1-T-AT2, both AT1 and AT2 were tested. In both cases, AT1 only was confirmed to inhibit the toxin's activity while we did not observe toxin inhibition activity with AT2.

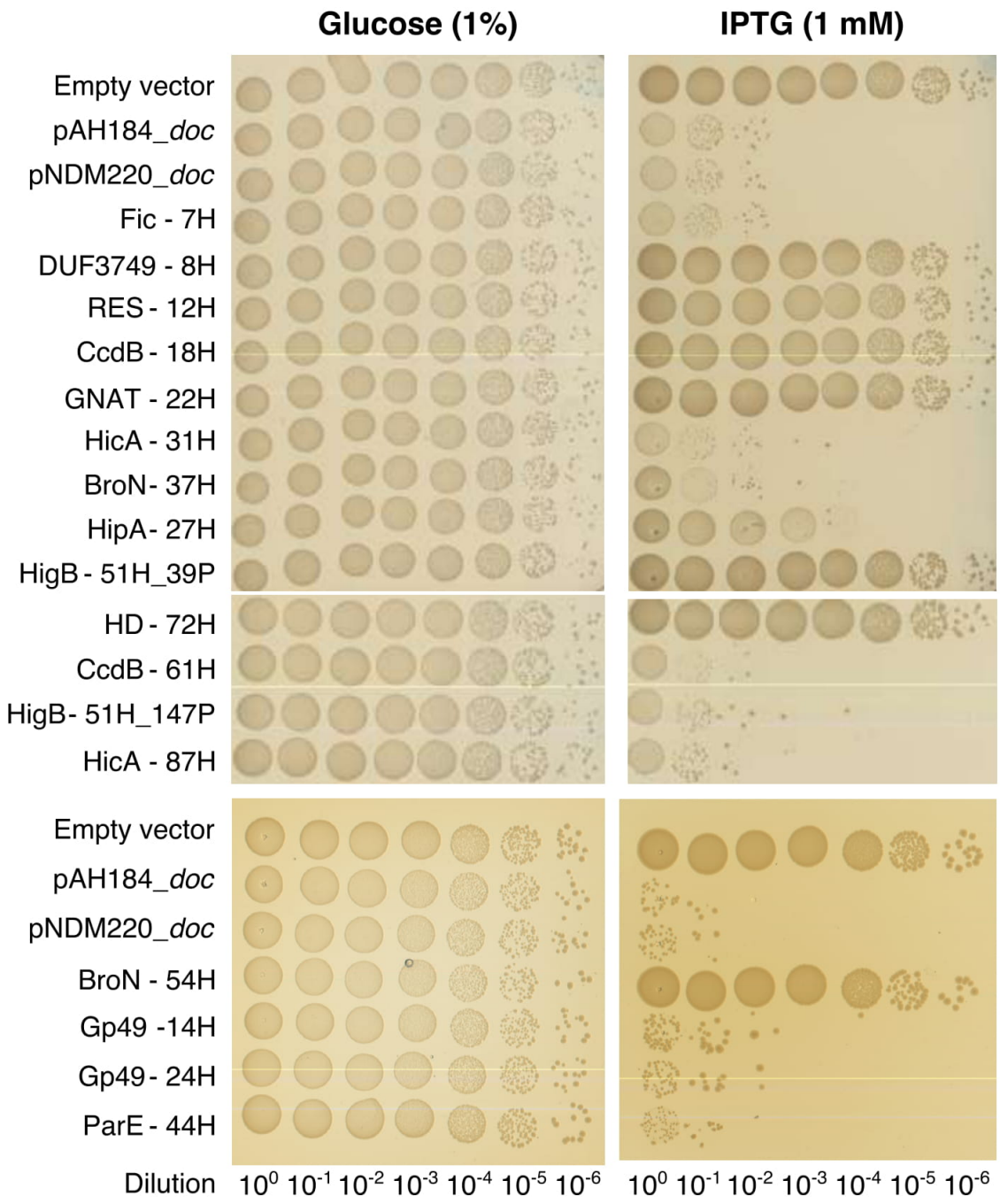

Figure 3.9: Phenotypic testing of selected toxins. *This work was conducted by Cinzia Fino, Matthew Dorman and Alexander Harms. LB agar plates were supplemented with 1mM IPTG for the induction of toxin Plac promoters. Overnight cultures were serially diluted (10^{-1} to 10^{-6}) in PBS containing the inducing supplements.*

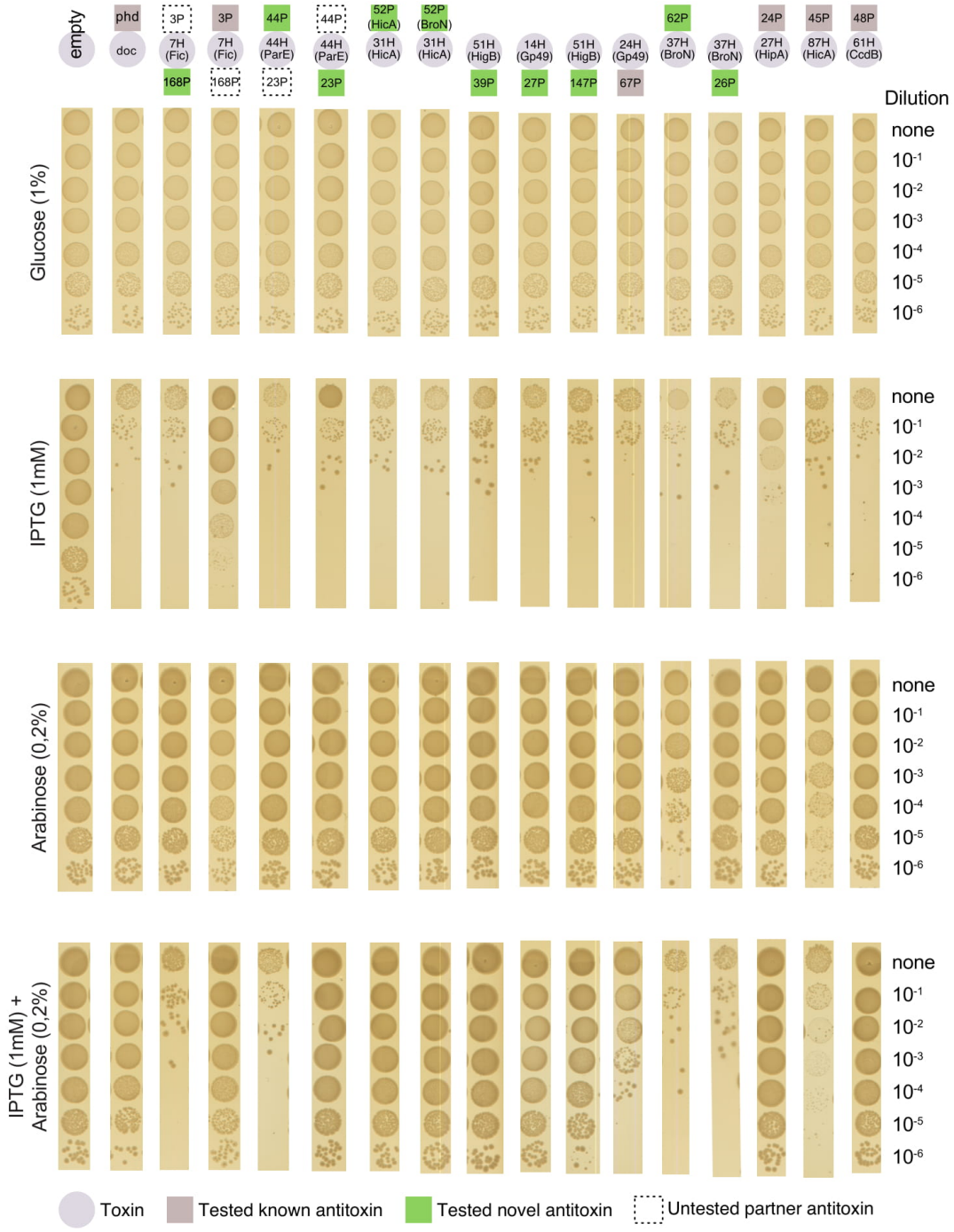

Figure 3.10: Phenotypic testing of predicted toxin-antitoxin combinations. *This work was conducted by Cinzia Fino, Matthew Dorman and Alexander Harms. Toxins in circles, antitoxins in squares. Tested novel antitoxins in green and tested known antitoxins in gray. For operon structures AT1-T-AT2, the untested partner antitoxin is in a dashed square. LB agar plates were supplemented with 1 mM IPTG for the induction of toxin Plac promoters' and 0.2% w/v of L-arabinose for the induction of antitoxin Para promoters'. Overnight cultures were serially diluted (10^{-1} to 10^{-6}) in PBS.*

Finally, these data revealed some more unexpected findings. In two cases the predicted antitoxins were themselves found to be toxic in our experimental system (45P, 62P) (Figure 3.10). One of these antitoxins is a well-described antitoxin with a HicB domain (62P). In addition, we confirmed both versions of antitoxin group 52P, associated with toxins from markedly different groups (31H (HicA) and 54H (BroN)), were able to counter toxin group 31H (Figures 3.10, 3.9; Table 3.2). Although the antitoxin group was linked to two different toxins and the two versions of the antitoxin shared only 83.2% amino acid identity, both versions inhibited the activity of this toxin. We were unable to confirm the toxicity of toxin group 54H (BroN) (Figure 3.9, Table 3.1), hence we could not confirm inhibition of this toxin group by these antitoxins. Finally, two variants of the toxin group 51H were tested (HigB); a shorter protein (53 aa) which was observed with antitoxin group 39P and a longer protein (103aa) observed with antitoxin group 147P. The C-terminus of the longer toxins was 83% identical to the shorter protein. The two antitoxins shared 71% amino-acid identity. We were only able to confirm the toxicity of the shorter 51H toxin. Nonetheless, we tested both antitoxins 39P and 147P with the shorter 51H toxin, and found that both antitoxins were functional and able to inhibit the toxin (Figure 3.10).

3.4.6 Orphan antitoxins are abundant in the dataset

We sought to determine whether the antitoxins of the TA pairs were also present on the *K. pneumoniae* species complex genomes as orphan genes uncoupled to a candidate toxin gene. The predicted antitoxin sequences were aligned against all the genomes and a total of 2,253 occurrences of orphan antitoxins belonging to 105 of the 233 antitoxin groups defined in this study were identified in the genomes (96 type II and 9 type IV) (Figure 3.11A, Appendix D). Of these, 25% were known antitoxins found in TADB or Interpro (26/105). For 80% (77/96) of type II and 89% (8/9) of type IV antitoxin groups, fewer than 26 orphan copies were identified in the entire genome collection, i.e. occurrences of unpaired antitoxins were rare and were found in fewer than 10% of genomes (Figure 3.11A). Conversely two antitoxin groups, containing the type II Fic and HipA toxin domains, were observed as unpaired in more than 80% of the genomes (>207 copies) across the species complex. In 35 of the 105 orphan antitoxin groups, orphans were detected in a species that was different to that of the original valid TA pair (Appendix D). For instance, antitoxin group 89P of the HipA toxin was originally identified in *K. quasipneumoniae*. However, orphan antitoxins were observed only in *K. variicola* (Figure 3.11A). Similarly, antitoxin group 115P belonging to a PemK-containing toxin was originally identified in *K. variicola*, but orphan antitoxins were observed in *K. quasipneumoniae* as well. Altogether there were no significant differences in the number of

orphan antitoxins per strain between the three species, with a median of nine orphans per strain across the three species (Figure 3.11B) (pairwise Wilcoxon rank sum test, FDR corrected).

Figure 3.11: Orphan antitoxins in K. pneumoniae genomes. *A Number of orphan antitoxins identified from each antitoxin group, coloured by K. pneumoniae complex species. The toxin Pfam profile of the toxin of the valid TA pair is in brackets. Antitoxin of type IV toxins are highlighted.* ***B*** *Orphan antitoxins per strain stratified by K. pneumoniae complex species.* ***C*** *Illustration of context analysis applied to each orphan antitoxin. The flanking sequences around each orphan antitoxin were compared to the flanking sequences of the valid TA pair. Each flank was classified according to whether or not it matched the sequence of the original TA pair.* ***D*** *Number of occurrences of orphan antitoxins classified by the similarity of their contexts to the valid TA pairs' .* ***E*** *Presence of a CDS in the orphan antitoxin's toxin-context.*

To assess the origin of orphan antitoxins, the upstream and downstream sequence surrounding the antitoxin were aligned with those found in valid TA pairs (Figure 3.11C) (See Section 3.3.6). 39% of the orphan antitoxins (879/2,253) shared the same toxin-context as the valid TA pair. Of these, 92% also shared the same non-toxin-context, indicating that they are in the same genetic context as the valid TA pairs from the same group (Figure 3.11D). 65% of orphans which did not share the toxin-context of the original TA pair (893/1374) did share the non-toxin context. In 20% of cases (470/2,253) neither the toxin-context or the non-toxin-context matched the valid TA pair, i.e. the orphan antitoxins were surrounded up- and downstream by unrelated sequences to any of the detected TA pairs.

To confirm whether these were truly orphan antitoxins, a CDS within the toxin-context was searched for that could function as the toxin. In 49% of orphans (1,107/2,253) a CDS within the context region was identified that does not contain a known toxin domain and could be a candidate for a novel toxin (Figure 3.11E). In 43% of cases (947/2,253) a toxin containing the original Pfam profile used in the search was found but the CDS was discarded due to the conservative structural requirements applied for a TA system (Figure 3.11E). These may be false negatives in the original analysis, or otherwise TAs which have diverged from the expected structure for a functional TA pair. In 8% of cases (171/2,253) the predicted antitoxin was truly orphan as a CDS longer than 50 aa could not be identified in the context region that may function as a toxin. In 1% of cases (28/2,253), the orphan antitoxin was close to the contig edge or proximate to a region with more than eight unknown nucleotides (N/X) and therefore the presence or absence of a toxin in its proximity could not be confirmed.

3.4.7 The association between toxins and antimicrobial resistance genes, virulence genes or plasmid replicons

Several of the sporadically distributed toxin groups were associated with clinically relevant AMR or virulence genes as well as plasmid replicons linked to the spread of AMR in *K. pneumoniae* and *E. coli* (Figures 3.12A,B, Fisher's exact test $p<0.01$, FDR corrected). These included 24H (Gp49) and 72H (HD) toxin groups which were significantly associated with multiple AMR genes, including those conferring resistance to aminoglycoside, amphenicol, sulfonamide, tetracycline and beta-lactams, with 13-29% of toxin genes found on the same contig as the respective AMR genes (Figure 3.12C). 100% and 30% of toxin CDSs' of toxin groups 24H and 72H respectively were on the same contig with an IncA/C plasmid replicon (Figure 3.12D). These contigs shared 99% (24H) and 97% (72H) sequence identity with the *K. pneumoniae* IncA/C-LS6 plasmid (JX442976), originally isolated from carbapenem-resistant *K. pneumoniae* [373], as well as AMR plasmids pNDM-KN (24H), pRMH760, pIMP-

PH114 and pR55 (72H) (Appendix D) [374–377]. Two toxin groups with a RES domain, 3H and 12H, were associated with multiple virulence genes (Figure 3.12B, Fisher's exact test p<0.01, FDR corrected) and one of these groups (3H) with the presence of an IncHI1B plasmid replicon. Contigs containing these two toxins showed over 99% sequence identity to *K. pneumoniae* virulence plasmids pK2044 and pLVPK (Appendix D) [378,379]. Five other toxin groups which were associated with AMR or virulence genes were also associated with the presence of plasmid replicons (Fisher's exact test p<0.01, FDR corrected) (see Figures 3.12A-C).

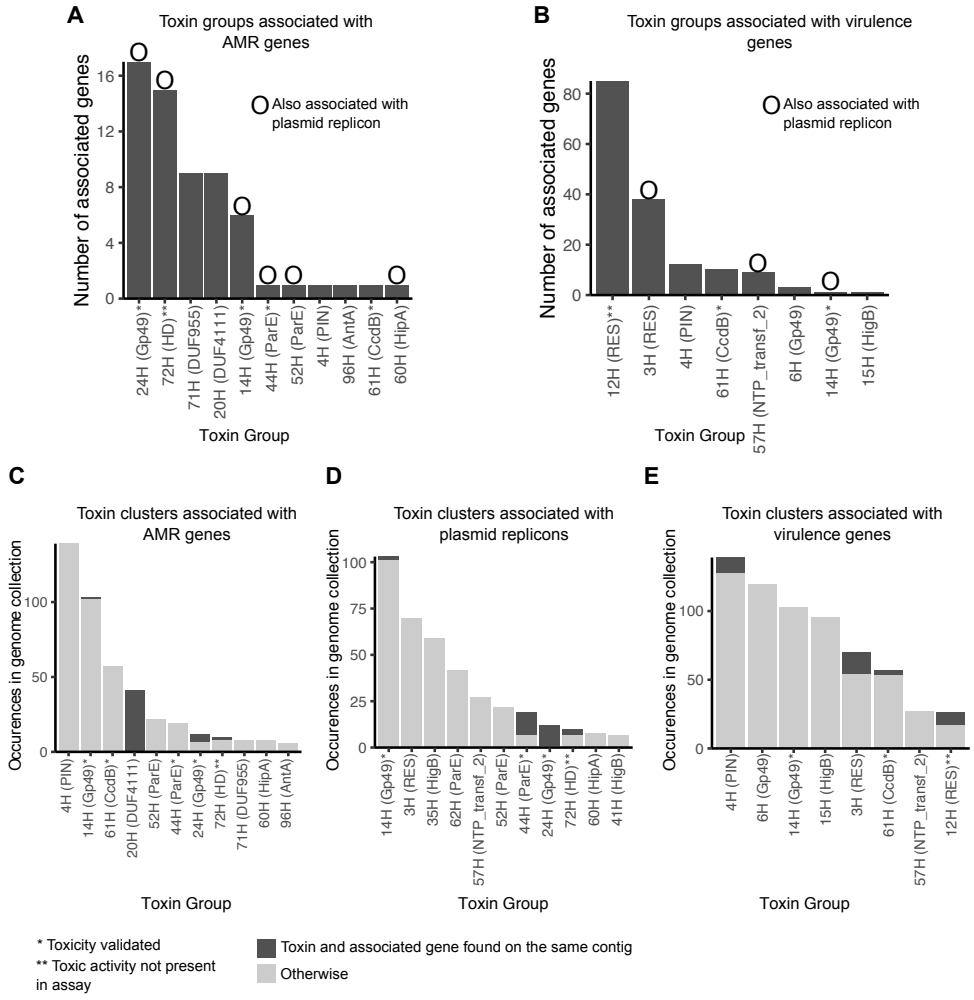

Figure 3.12: Toxin groups associated with AMR genes, virulence genes and plasmid replicons. *Number of unique AMR **(A)** and virulence **(B)** genes associated with each of the toxin groups. Circles above bars indicate the toxin group was also associated with the presence of a plasmid replicon. **C-E** Number of occurrences of toxins in the genome collection, for the toxin groups associated with AMR genes (**C**), plasmid replicons (**D**) and virulence genes (**E**). An occurrence of a toxin is coloured in dark if it was observed on the same contig with one or more of the associated genes, light otherwise.*

3.5 Discussion

In this chapter was presented a systematic in-depth analysis of the diversity and evolution of TA systems in a large collection of a clinically important member of the *Enterobacteriaceae,* the *K. pneumoniae* species complex. TA systems are highly prevalent in the species complex*,* however, the underlying processes of the evolution of TA systems are likely to be context-dependent. The toxins of these TA systems could be classified based on their distribution patterns as ubiquitous, species associated, sporadically distributed (often with associations to clinically important genes) or rare. The evolution of ubiquitous toxins is likely vertically inherited, as higher nucleotide identity was observed between toxins of the same species than between species. The same effect was not observed for the sporadic toxins, suggesting that some TA systems are more mobile than others. Importantly, the classification presented in this study was based on the dataset used, which was aimed to capture the diversity of the *K. pneumoniae* species complex*.* It is possible that further sampling of under-represented lineages would increase power and refine the classification.

The pairing of antitoxin to toxin is not fixed; for each toxin a range of candidate antitoxins were identified and found in different arrangements, putatively able to inhibit the same toxin. Sampling of more genomes would lead to a large diversity in antitoxins relative to toxins, suggesting the potential number of interactions between toxins and antitoxins is large. Notably, some toxins were more stably coupled to a single antitoxin and observed in a single arrangement, while other toxins were observed with a wide range of antitoxins and operon arrangements. This highlights that the co-evolution between toxin and antitoxin is dependent on the system and context. This has functional implications as the antitoxin and its interaction with the toxin can affect the functioning of the TA system ([380]. Some antitoxins play a role in the regulation of the TA module as the toxin-antitoxin pair regulate the expression of the TA operon [346,371]. Furthermore, the interaction of the toxin with the antitoxin will determine the specificity of the inhibition and therefore would affect the dynamics of both activation and deactivation of the TA operon. Finally, antitoxin instability is often the result of degradation by proteases [346], therefore the inhibition of an antitoxin in response to stress can depend on the antitoxin sequence as it would determine the specificity of interaction with proteins that lead to its degradation [381].

Even more, a number of toxin or antitoxin groups were observed as specific to a species, i.e. a toxin-antitoxin pairing was observed only in one particular genetic background. This suggests it may be beneficial to possess a specific toxin-antitoxin pair under one genetic background compared to another.

Altogether 76% of the identified candidate antitoxins were novel and not identified in the existing toxin-antitoxin database TADB or Interpro [318,319,382]. Furthermore, there was additional sequence diversity within each antitoxin group that we found. These results emphasise the potential large diversity of antitoxins that could inhibit these toxins and our lack of knowledge of the complete range and diversity of these systems.

Using an *E. coli* model system, the toxicity of 10 of 17 tested toxins was confirmed (~59%) and the inhibitory activity of 10 of 14 tested antitoxins (~71%). Nine of the tested antitoxins are novel and we were able to confirm the inhibition of five of them. We also found candidate antitoxins downstream of the toxin, and confirmed the inhibitory activity of three of them, highlighting exceptions to the common setup in which the antitoxin is encoded upstream of the toxin. These results could form the basis of future studies investigating how different autoregulatory principles enabled by upstream or downstream antitoxins might affect the biology of a TA system. While some of these candidate antitoxins could be false predictions, the observation of known or confirmed antitoxins both upstream and downstream to toxins suggests we cannot rule out any antitoxin candidate. Importantly, a negative result in our assays does not rule out toxic or inhibitory activity of these proteins, but rather could be the result of confounding effects in our assays for example biological differences between *E. coli* K-12 and *K. pneumoniae*, lack of protein expression or incorrect folding in the heterologous host. Furthermore, our assays do not indicate whether these systems are expressed in the host bacterium or whether they have a physiological role in the host cell.

There is an abundance of orphan antitoxins present in the population which are unpaired to a functional toxin. These include a number of the antitoxins we expressed and were able to confirm their inhibitory activity (92 orphan copies of 39P, 17 orphan copies of 24P and 45 orphan copies of 45P, Figure 3.11). Sources of orphan antitoxins may be degrading TA pairs that are in different genetic locations, older degraded TA systems or otherwise, these could be candidates for new toxins which share the same antitoxin as we have identified. Alternatively, some orphan antitoxins may be paired to a known toxin but were discarded in our analysis due to the conservative structural criteria we defined for a TA system, suggesting that the prevalence of TA system in the *K. pneumoniae* species complex presented here may be under-estimated.

These orphan antitoxins may be serving a new purpose. For example, they may serve as anti-addiction modules, preventing the fixation of plasmids or other MGEs [383]. They may be interacting with the toxins of active TA pairs and affecting their function. Alternatively, they

could also be conserved as remnants of a degraded TA locus that have acquired functions in transcriptional regulation of other genes in the genome [384].

The importance of this type of analysis is not limited to TA systems, and presents general trends to distinguish between groups of genes of other gene systems. Pan-genome analysis of bacterial datasets is often focused on the description of core compared to accessory genes without focusing on the precise details within these two categories. Here we showed a more refined description of genes based on their distribution across the *K. pneumoniae* population and in the context of linkage to other genes. This finer grained analysis can be applied in other settings and lead to novel, highly relevant insights on evolutionary dynamics of poorly understood genetic elements.