2 SLING: A tool to Search for LINked Genes in bacterial datasets

This chapter is a modified version of the published paper "SLING: a tool to search for linked genes in bacterial datasets" [311]. Alexander Harms, Cinzia Fino, Leopold Parts, Kenn Gerdes, Eva Heinz and Nicholas Robert Thomson contributed to the research of the original publication. All final language is my own.

2.1 Introduction

Operons or functionally linked gene arrays represent the most basic unit of transcriptional organization in prokaryotic genomes [312]. Genes involved in the same process or pathway are encoded in a single block, and transcribed under the same regulation [312]. Identifying homologues for a single gene is a difficult task that has been tackled using many methods, as was described in Section 1.4.1.2. The identification of two genes or more which are physically linked to each other further complicates the search. This is because the structure of operons and gene arrays with similar functions can vary substantially across isolates and species. The order of the genes is often changed, and individual genes may be lost or gained [313–315].

TA systems are an example of a simple two-gene operon and were presented in Section 1.3. Databases have been constructed which enable the search for TA systems using simple homology based search tools such as Blast+ [227,251,316–321]. The most well-curated and accessible database is the TA database, TADB [318,319]. However, a homology-based search does not always verify whether the identified genes represent intact CDSs or whether the toxin and the antitoxin are adjacent, meaning further downstream manipulations are required. Two tools have been published which allow for a direct search of the toxin and the antitoxin: RASTA and TAfinder, the TA search tool provided within TADB [318,322]. However, both of these tools are provided in an online interface which is not scalable when examining these systems on larger scales. Even more, RASTA, which was published over a decade ago, no longer in service. Furthermore, they do not allow the addition of custom sequences or domains in the search [318,323,324]. This limits the search, and the quality and relevance of the annotation is determined by the quality of the database. Users have to rely on updates to obtain the most up to date results.

Many other clinically important gene systems are encoded in operons; all secretion systems [323,325], CRISPR-cas systems [315,326], Resistance Nodulation Division (RND) efflux

pumps [327], and more follow this organization. For these more complicated operon structures, sophisticated methods have been developed for their annotation [318,322–324,328]. These tools are restricted to the specific operon which is being investigated as they rely on previously defined structures and sequences, or require reprogramming for identification of new genetic structures.

With the growing availability of large datasets for the surveillance of important pathogens [9,329,330], there is a need for a single flexible framework to annotate clinically relevant gene arrays across a range of isolates and examine their diversity. While a level of specificity will always be required to define the search of a specific operon, there is room to develop generic methods which could search for a range of operons with only a few input requirements from the user.

2.2 Aims

The aim of this chapter was to develop a tool to search for and group operons in large bacterial datasets. In many operons or gene arrays, there is a single conserved gene which is always present together with its neighbours in a rule-defined proximity and orientation. This property provides the potential to capture the diversity of the gene array based on the diversity of the single conserved gene and its neighbours. The precise aims of this chapter were:

- Define the SLING pipeline, a tool to Search for LINked Genes
- Construct the required settings to search for TA systems, and apply these on a collection of *E. coli* isolates.
- Construct the required settings to search for RND Efflux Pumps and apply these on a collection of *E. coli* isolates.

2.3 Methods

2.3.1 SLING specifications

SLING was implemented in Python (2.7) and is available to download from https://github.com/ghoresh11/sling. The steps of the SLING pipeline are detailed in Section 2.4.1 and in Figure 2.1.

Genome preparation Complete genomes or assembled contigs in FASTA format were six-frame translated using Biopython v1.68 [331]. By default, translation is performed using the standard codon table and the permitted start codons are [ATG, TTG, GTG]. SLING will search

for the longest CDS beginning with ATG, if it is not found it will search for the longest CDS beginning with TTG and finally GTG. Annotation files of the provided genomes in GFF format can also be provided.

Searching HMMER (v3.1b2) [296] was used to search all CDSs for the profiles of the primary gene provided by the user. The cut off used for a CDS to be considered a 'hit' for downstream analysis is a HMMER bit score of the overall sequence/profile comparison of at least 20. The cutoff was chosen based on the scores of toxin HMM profiles in known toxin sequences downloaded from TADB [318,319].

Filtering 'Partner' genes were searched in proximity to the hits according to structure requirements provided by the user. The structure requirements include the orientation of the partner gene relative to the conserved gene (upstream, downstream, or both for a three-component array), the minimum and maximum length of the conserved gene, the minimum and maximum lengths of the partner genes (upstream and downstream if applicable), and the limitations on the location of the partner gene relative to the conserved gene (maximum overlap and distance). If no partner is found under the given requirements, the hit is discarded. For the built-in HMM collections presented in this thesis, these requirements are provided by SLING; however, the default values can easily be overridden. Partner genes which have eight or more consecutive unknown nucleotides (Xs or Ns) are removed at this stage and not considered by SLING.

Profile-specific length requirements. The user can provide SLING with a file containing the expected length of proteins of each of the profiles in the HMM collection, and a limit on the maximum permitted difference between a hit's length and its expected length. This is useful when scanning for multiple profiles of conserved proteins that have versatile expected lengths.

Grouping Sequence similarity networks (SSN) are constructed for all the hits and the partners identified using protein-protein BLAST+ (v2.7) [285]. When using an orientation requirement of "either", SLING will treat upstream and downstream partners the same to form a single SSN. When using "both", SLING will generate an SSN for the upstream partners and the downstream partners separately.

Each node in an SSN is either a hit or partner sequence. An edge is drawn between two hit nodes or two partner nodes only if they meet the minimum requirements of sequence similarity as provided by the user for the BLAST output. The default requirements applied for the results

in this paper are an e-value of 0.01 and a percent identity of 30. All sequences found in the same connected component in the SSN are considered to be in the same hit/partner group.

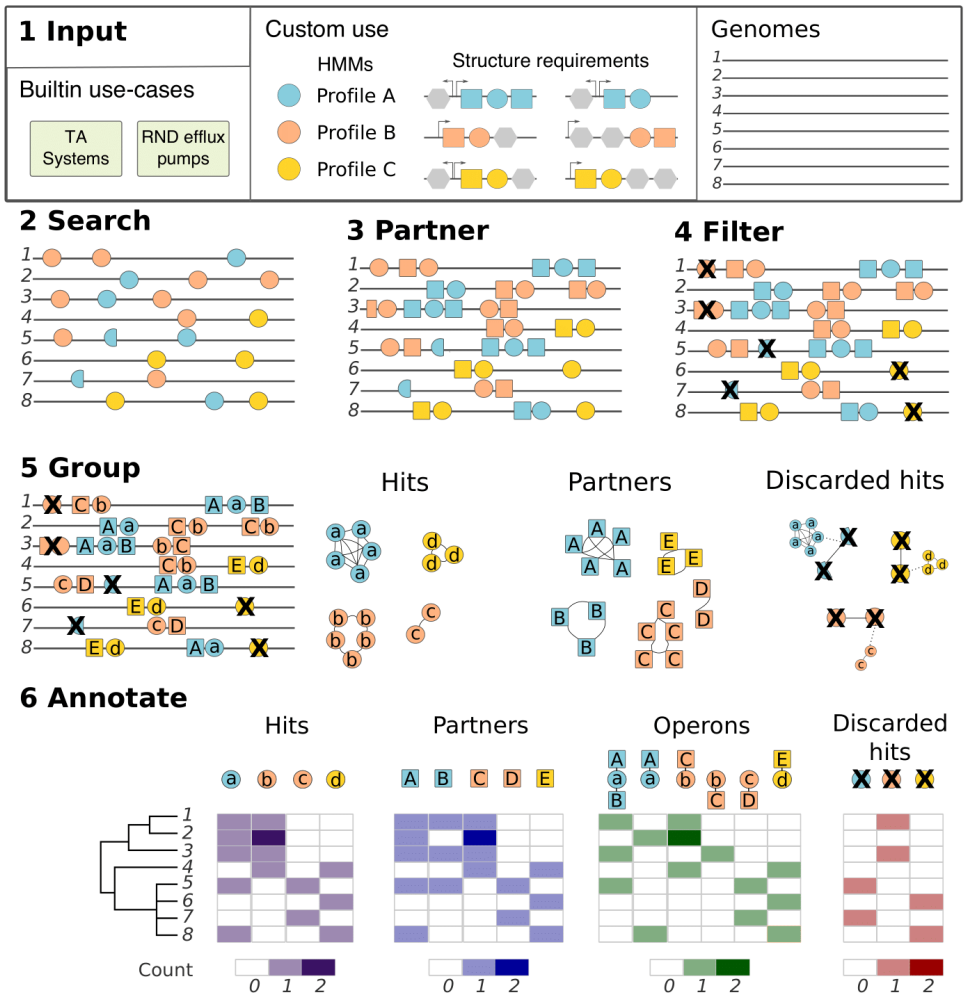

Figure 2.1: Overview of the SLING pipeline. *(1) SLING input. The user may use one of the built-in cases or otherwise provide SLING with a collection of HMM profiles and structural requirements. The structural requirements presented provide a simple example of gene arrays with multiple possible structures (top left). Grey octagons represent variable genes. Circles represent conserved genes each with a matching HMM profile represented by a unique colour which are used in the SLING search. Squares represent the partner genes consistently found in a rule-defined proximity to the conserved gene. (2) HMM profile hits are found in the input genomes. (3) Partner genes are located. (4) Partner genes are filtered based on the given structural requirements. (5) Hits, partners and discarded hits are grouped (alphabetic labelling) using sequence similarity networks. Discarded hits are mapped back to the accepted hits. (6) SLING outputs can be loaded into ITOL for visualisation of results. The phylogenetic tree must be provided for visualisation.*

Reporting discarded HMM matches The discarded hit sequences are grouped in an SSN as described above. Each connected component in this network is then mapped back to the clusters in the hits network and the discarded hit clusters are labelled according to their equivalent hit cluster.

2.3.2 Strains and phylogenetic analysis

The core gene phylogeny of 91 EPEC strains taken from [115] (See Section 1.1.2.3) was inferred from a core gene alignment generated using Roary [305], and a maximum likelihood tree from the informative single nucleotide polymorphisms (SNPs), chosen using SNP-sites [332] (v2.3.2), was constructed using RAxML (v8.2.8) [282] with 100 bootstrap replicates.

2.4 Results

2.4.1 SLING overview

SLING is a command line tool which requires a collection of assembled genomes (contigs or complete), HMMs representing a conserved gene within the gene array of interest and optional structural requirements as input (Figure 2.1). Each HMM profile is used to search the genomes for the presence or absence of the primary gene. If the gene is detected, referred to as a 'hit', SLING attempts to identify the partner protein CDSs proximal to it. The results are filtered to match the provided structural requirements. These include the distance between the partner and hit, their permitted lengths and the orientation of the 'partner' gene relative to the conserved gene. If the structural requirements are unknown, SLING will search for the closest neighbouring genes with no limitations. Hits, partners and discarded hits are grouped using SSNs. Finally, SLING reports the number of occurrences of each hit group, partner group, complete array group and discarded hit group found in each genome. These can easily be loaded into statistical analysis tools or into ITOL [333], an online tool for display and management of phylogenetic trees, creating an immediate interface for the user to examine the distribution across large datasets. SLING is available to download from https://github.com/ghoresh11/sling. Full details are provided in Section 2.3.1 and in the package wiki (https://github.com/ghoresh11/sling/wiki).

2.4.2 TA systems search

SLING can be used to search for simple two-component operons, such as TA systems. As SLING is based on a CDS search, the focus is on type II TA systems where cognate antitoxin is a protein which inhibits the toxin through direct interactions [238] (Figure 1.6). For a

complete introduction on TA systems, refer to the Section 1.3 of the Introduction. Type II systems are well studied and their structure is generally known; the antitoxin and toxin genes are transcriptionally coupled with well defined rules describing the gene orientations and distance separating them [238,316]. Moreover, TADB, which has an extensive database of type II TA systems, was available as a resource to benchmark the approach [318,319] . Following the same set of rules, type IV systems were also included in which the antitoxin is also a protein which inhibits the toxin's activity via the toxin's target [334]. Only a few type IV systems have been described so far, and appear to be rare compared to the abundant type II TA systems [334].

2.4.2.1 Construction of profile HMM library and structural requirements

To generate a collection of toxin HMM profiles, used as the primary gene in SLING, type II and type IV toxin sequences were retrieved from the web based resource for TA loci, TADB [319] and were supplemented by additional toxin sequences based on a literature search. All the toxin sequences were scanned against the Pfam protein domain database (v30.0) with HMMER (v3.1b2) to identify known toxin domains, obtaining an initial set of 153 putative HMM profiles [296,298]. These HMM profiles were manually curated to remove antitoxin domains and domains of non-protein-based TA systems which were not the subject of this investigation. Additionally, HMM profiles which had fewer than five hits were removed for further analysis unless they were a domain of a well described toxin.

A test dataset of 33 *K. pneumoniae* genomes and plasmids taken from [335] was scanned with the remaining HMM profiles. This dataset was used in order to characterise the Pfam profiles on a small collection of genomes. For each profile, the total number of HMMER hits were counted across the 33 genomes and their average length was compared to the length of the toxins containing the same profile on TADB (Figure 2.2A,B). This enabled the identification and removal of Pfam profiles which had many hits of the expected length of a toxin that do not always represent a true toxin. Keeping such profiles in the TA search would lead to high false discovery rate. For instance, the Acetyltransf domains often had a high number of hits within the expected length of a toxin and were removed (Figure 2.2B,C). Other profiles, like DUF294 and NTP_transf_2 did not have many hits, however, they did show high variability in their length relative to the lengths of the toxins containing them on TADB. For these toxins, their profiles were kept in the search and an option to apply a profile-specific-length limitation within SLING was added. Thus, only hits which were up to 100 aa longer or shorter than the average toxin length were accepted for downstream steps (Figure 2.2D). Finally, most profiles showed both a low hit count as well as fell within the range of expected lengths (Figure 2.2B,E). The final collection, following this curation step, consisted of 54 toxin profiles.

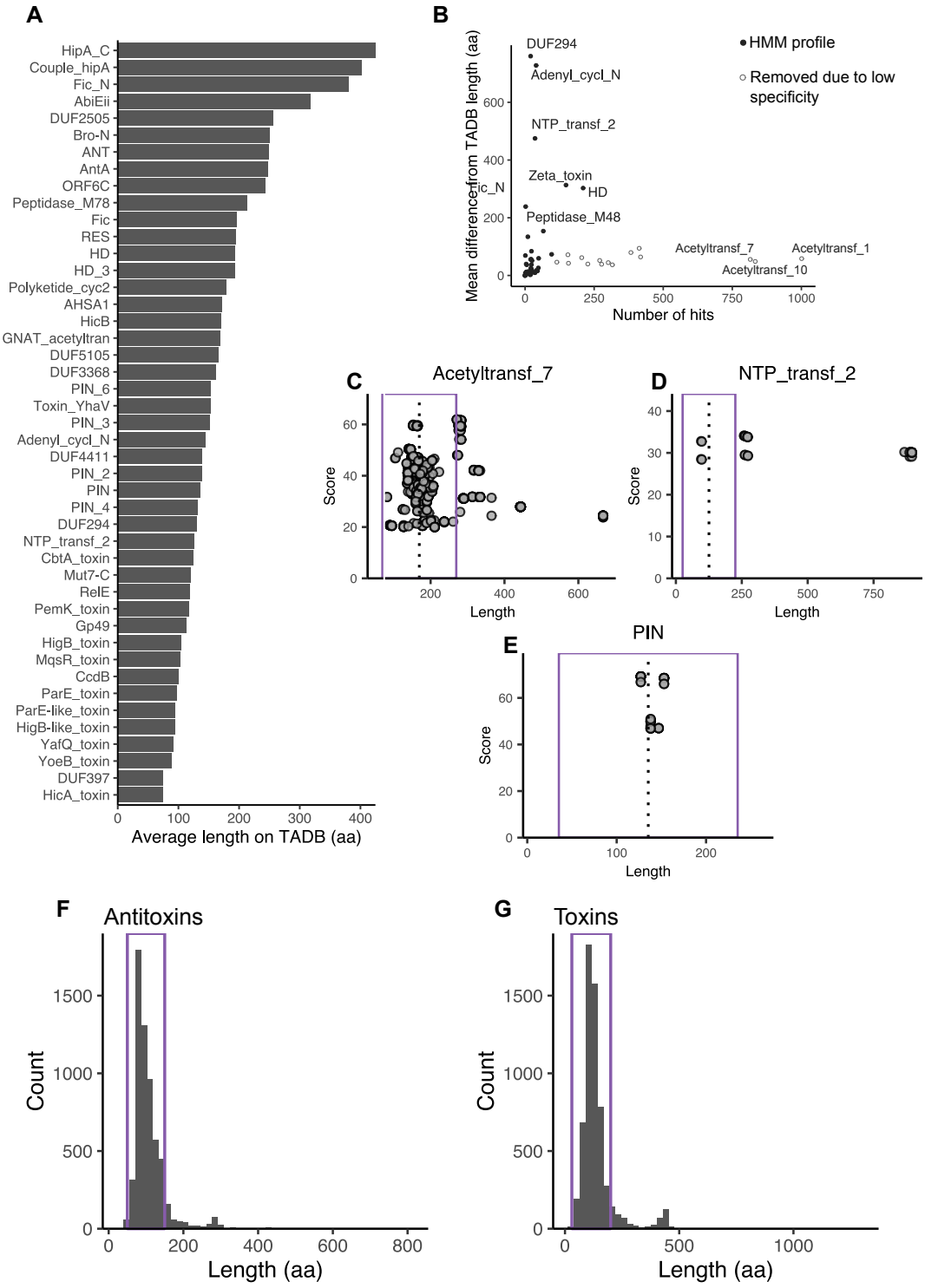

Figure 2.2: Defining the HMM collection and structural requirements for toxins. A *Mean length of toxin sequences in TADB* ***[318,319]*** *containing each of the HMM profiles.* ***B*** *Number of hits in 33* Klebsiella *genomes relative to the mean difference of those hits in protein length relative to the profiles' mean length as found on TADB (presented in* ***A)****. Empty dots are profiles which were removed due to low specificity as there were many hits which differed significantly in length relative to the length of the protein in TADB.* ***C-E*** *Length of all hits in 33*

Klebsiella *genomes relative to their HMMer bit-score. Dotted line represents the mean length of the profile in TADB (as presented in **A**). Purple rectangle represents the length cut-off defined in SLING for an ORF to be considered a valid toxin. **C** Example of low specificity HMM profile which has been removed. **D** Example of HMM profile with large length range, but with high specificity for ORFs within the expected length range. **E** Known toxin domain with small length-range and number of hits. **F, G** Length distribution of all antitoxins (**F**) and toxins (**G**) downloaded from TADB. Purple rectangles represent the length cut-offs defined in SLING.*

The length distributions of the toxin and antitoxin sequences downloaded from TADB were plotted to define the length requirements. Over 90% of antitoxins were between 50 and 150 aa long; therefore, these were used as the relevant cut-offs (Figure 2.2F). The permitted length of proteins containing toxin profiles which were present in TADB was determined based on their mean length in TADB (detailed above). Some toxin profiles were taken from a literature search and thus were not present in TADB and an average length was unavailable. For these, a minimum length cut-off of 30 aa and maximum length cut-off of 200 aa were chosen as these covered over 90% of toxin sequences in TADB (Figure 2.2G).

Table 2.1 Search parameters used in SLING

	Default	TA systems	RND efflux pumps
Order	either	either	upstream
Minimum hit length (aa)	1	30	700
Maximum hit length (aa)	10000000	200	1500
Minimum downstream length (aa)	1	50	NA
Maximum downstream length (aa)	10000000	150	NA
Minimum upstream length (aa)	1	50	100
Maximum upstream length (aa)	10000000	150	1000
Maximum distance between hit and partner (bp)	10000000	50	20
Maximum overlap between hit and partner (bp)	300	20	500
Maximum difference from average length (if given) (aa)	10000000	100	200

Finally, a distance of up to 50 bp and an overlap of at most 20 bp were permitted between the toxin and antitoxin genes. The orientation requirement was set based on the knowledge that the partner gene, i.e. the antitoxin, can be either upstream or downstream of the toxin gene (Table 2.1) [316].

2.4.2.2 The process for setting up a TA search are applicable to other operons

A similar process can be applied to construct the HMM profile libraries of other genes and to define the structural parameters. Another example will be presented in Section 2.4.3.1 and the general approach is summarised in Figure 2.3. HMM profiles can also be generated directly from an MSA of a collection of genes using HMMER [296]. Finally, if the structural requirements are unknown, SLING provides default parameters for a flexible search which will identify the closest partner genes proximate to the primary gene (Table 2.1).

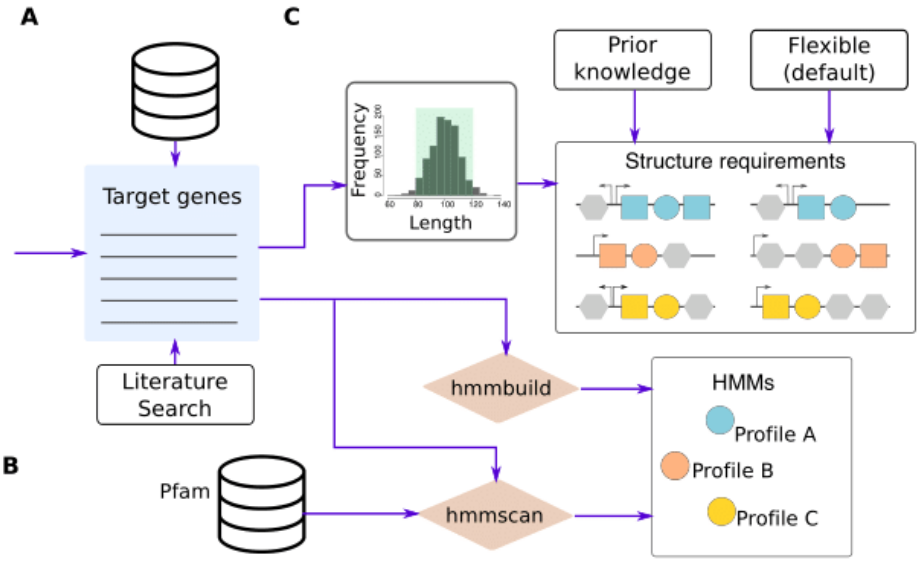

Figure 2.3: General construction of HMM profiles and structural requirements for SLING input. A *A collection of known target genes is required, taken from existing databases (toxins; TADB, RND pumps; Uniprot), a literature search or other sources.* ***B*** *HMM profiles can be generated directly from an MSA of the target sequences using HMMER [296]* hmmbuild *or can be scanned by HMMER* hmmscan *against existing HMM profile databases, for instance, Pfam [298].* ***C*** *Structural requirements can be inferred from the target gene sequences, known from prior knowledge or otherwise, flexible using SLING's default parameters.*

2.4.3.3 Benchmark on *E. coli* K-12

SLING identifies new and known TA systems in *E. coli* K-12.

SLING was used to search *E. coli* K-12 strain MG1655 (NC_000913.3) for TA systems. SLING identified 23 TA systems in total (Figure 2.4B). These results were compared to the *E. coli* K-12 strain MG1655 TA systems in TADB and those predicted by TAfinder using the same parameters used in SLING [318,319]. Nine of the 23 systems were identified by all three methods. TADB missed five TA predictions which were identified by the other two methods, whereas TAfinder missed one. A single system, identified by TADB, was missed by both SLING and TAfinder, the rnlAB system. The RnlA toxin has a length of 397 aa, beyond the maximum length threshold of 200 aa for a toxin applied in our implementation.

SLING identified eight TA systems which were not predicted by TADB or TAfinder. Of these, four have been predicted in the past to be TA systems; the YkfI-YafW system [334,336], the GnsAB TA system [337], the RatAB system [338] and the YdaST system [339]. Four more predictions have not been previously described as TA systems and are candidates for further investigation. One contains an HD domain, two contain a GNAT domain and the last contains a YdaT toxin domain, consistent with their proposed function.

TADB and TAfinder identified TA systems that were not identified by SLING. Thirteen of the TADB results belonged to TA system classes that were not investigated in this study. An additional two toxins were predicted which, using HMMER, did not contain any described toxin profile used by SLING. Finally, TAfinder predicted three TA systems which we attempted to retrieve from the reference genome but were unable to identify complete CDSs at the relevant locus.

2.4.2.4 Application on EPEC collection

To search for TA systems in a diverse set of related bacteria SLING was applied with the settings described for TA search on a collection of 70 EPEC isolate genomes taken from [115], supplemented by an additional 21 commonly studied *E. coli* reference strains (taken from [115]). The EPEC isolates were collected from children presenting with diarrhoea from seven centres in Africa and Asia [115].

SLING identified a total of 94 different TA operons in the complete *E. coli* collection built of 44 toxin (hit) clusters and 80 antitoxin (partner) clusters. SLING generated an output of the absence and presence of these systems across the dataset that can be loaded into a statistical learning tool, enabling to look for association with the metadata and view in ITOL. Below are examples of three toxins which are presented to illustrate the type of visualisation, analysis and interpretation that can be accomplished using SLING (Figure 2.4C).

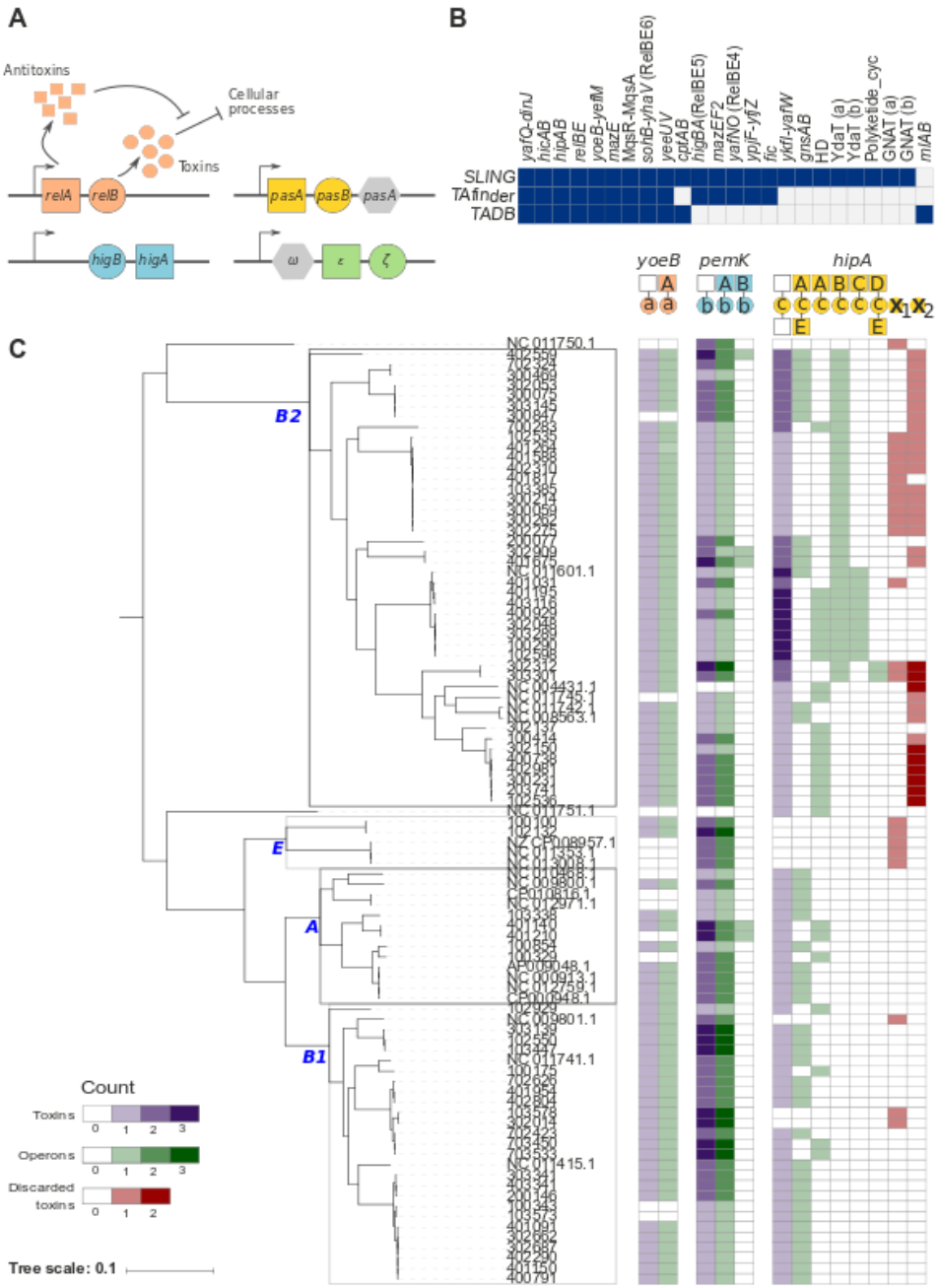

Figure 2.4: Identification of TA systems using SLING. A *Possible operon structures of TA operons. Each toxin has a unique HMM profile, represented by a different colour.* ***B*** *Identification of TA systems in* E. coli *K-12 using SLING, TADB and TAfinder. Prediction of a TA operon by a method is represented in a dark blue square. Novel TAs predicted using SLING named by the Pfam profile by which they were identified.* ***C*** *Description of the diversity of three toxins and their cognate antitoxins in the* E. coli *collection. Darker squares represent presence of a toxin or operon in an isolate.*

YoeB toxin presents low antitoxin repertoire, with low evidence of gene loss/gain. The YoeB profile containing toxin was always identified as partnered to the same antitoxin. This TA pair was ubiquitous, present across all phylogroups. In addition, there was no evidence of duplication events, with a single copy of the operon identified in each isolate. *yoeB* was never found as an orphan toxin, however there were examples of loss or gain of the whole operon in nine locations in the phylogeny, i.e. the antitoxin was never lost on its own. This observation strengthens the hypothesis that this protein serves as a toxin in a TA system.

PemK toxin presents medium antitoxin repertoire, with high evidence of gene loss/gain. The second toxin (Figure 2.4C), containing a PemK profile, showed diversity in its antitoxin repertoire: it was found with two different antitoxins: A and B. Most copies of this toxin were observed with one of the antitoxins (A; 97%), which was present across all the phylogroups. For this operon, there was a strong indication of gain events followed by fixation and vertical propagation; a subclade with a copy number of *n* was often found within a clade with copy number *n-1*. This phenomenon occurred independently multiple times in the phylogeny. The pervasiveness of this operon can either allude to its importance, or otherwise, suggests it is successful at spreading in the population and persisting. The second operon (B) was rare and found only in five isolates in a single copy. It was most likely acquired in three independent events. Finally, like *yoeB* toxin, this toxin was always found partnered to an antitoxin.

HipA toxin presents a high antitoxin repertoire, with low evidence of gain/loss of the same genes. The final toxin (Figure 2.4C), containing a HipA profile, presents a higher diversity in its antitoxin repertoire with five candidate antitoxins. Four of these antitoxins (A-D) are upstream to the toxin, whereas the last antitoxin (E) was found downstream to the toxin and was always present with one of the upstream antitoxins.

Looking at their phylogenetic distribution, although many of the isolates have more than one copy of the *hipA* toxin, it was apparent that within one genome each individual toxin gene was partnered with a different antitoxin. The majority of toxin genes were linked to antitoxin A (62%), which together were present across all phylogroups (Figure 2.4C). The three other antitoxins (B, C and D) are lineage specific and were only present in phylogroup B2. Interestingly, all isolates with antitoxins C or D also had antitoxin B.

Although *hipA* is a well described toxin, we observed multiple cases in which SLING filtered the predicted toxin gene out due to deviations from the expected operon structure of a TA system (Figure 2.4C). These genes were marked as discarded by SLING as a result of this. However, analysis of these discarded toxins showed that they formed two separate sequence

clusters: X_1 and X_2. All the X_1 toxins coincided with isolates which were missing the A antitoxin. As for X_2, all the discarded toxins were within phylogroup B2, coinciding with isolates which were missing antitoxins B and C.

2.4.3 RND efflux pumps search

Efflux pumps play an important role in multidrug resistance as they confer a mechanism for the efflux of antibiotics [340]. One example of this are the RND family of membrane transporters found in Gram-negative bacteria [327,341]. RND family pumps consist of three components: an outer membrane protein (OMP), a periplasmic fusion protein (MFP) and an RND pump (Figure 2.5A). In most cases, the MFP and RND components are found in an operon, whereas the OMP is located in a different location [327]. RND efflux pump operons, unlike TA systems, are complex operons which often include a large range of genes often found in different orders and orientations [327]. However, these operons always contain an RND efflux pump protein which is highly conserved and, in most instances, the MFP is located upstream of it and transcriptionally coupled to it [327]. This property makes these operons relevant for a search using SLING by setting the RND protein as the primary gene and applying flexible structure requirements on the partner gene.

2.4.3.1 Construction of profile HMM library and structural requirements

3,325 RND efflux pump sequences were downloaded (on 07.11.17) from Uniprot [342] by searching for the name of 26 known RND pump genes (Figure 2.6A) [343]. The sequences originated from 295 different genera. Sequences were clustered using cd-hit (v4.7) to remove redundant sequences which share 90% identity [344]. The remaining 1,242 sequences were searched using HMMER (v3.1b2) against the Pfam database (v30.0) to identify known RND pump domains [296,298] (Figure 2.3B). A total of 29 Pfam profiles were identified in these sequences, of which a single profile, ACR_tran (PF00873), was present in over 99% of the sequences and thus was chosen to represent all RND pumps.

The length distribution of the above mentioned RND pump proteins were plotted (Figure 2.6B). A minimum length of 700 aa long and maximum length of 1500 aa long were chosen for the RND pump protein, covering over 94% of the downloaded sequences. For the partner gene, 23,133 MFP sequences were downloaded (on 07.11.17) from Uniprot [342] by a keyword search. The length distribution of these proteins was plotted and a minimum length of 100 aa and maximum length of 1000 aa were chosen as flexible requirements for different partner genes as these thresholds cover the length of over 99% of membrane fusion proteins

downloaded [342] (Figure 2.6C). Finally, a maximum of 500 bp distance between the partner and the RND pump, and at most 20 bp overlap were allowed (Table 2.1).

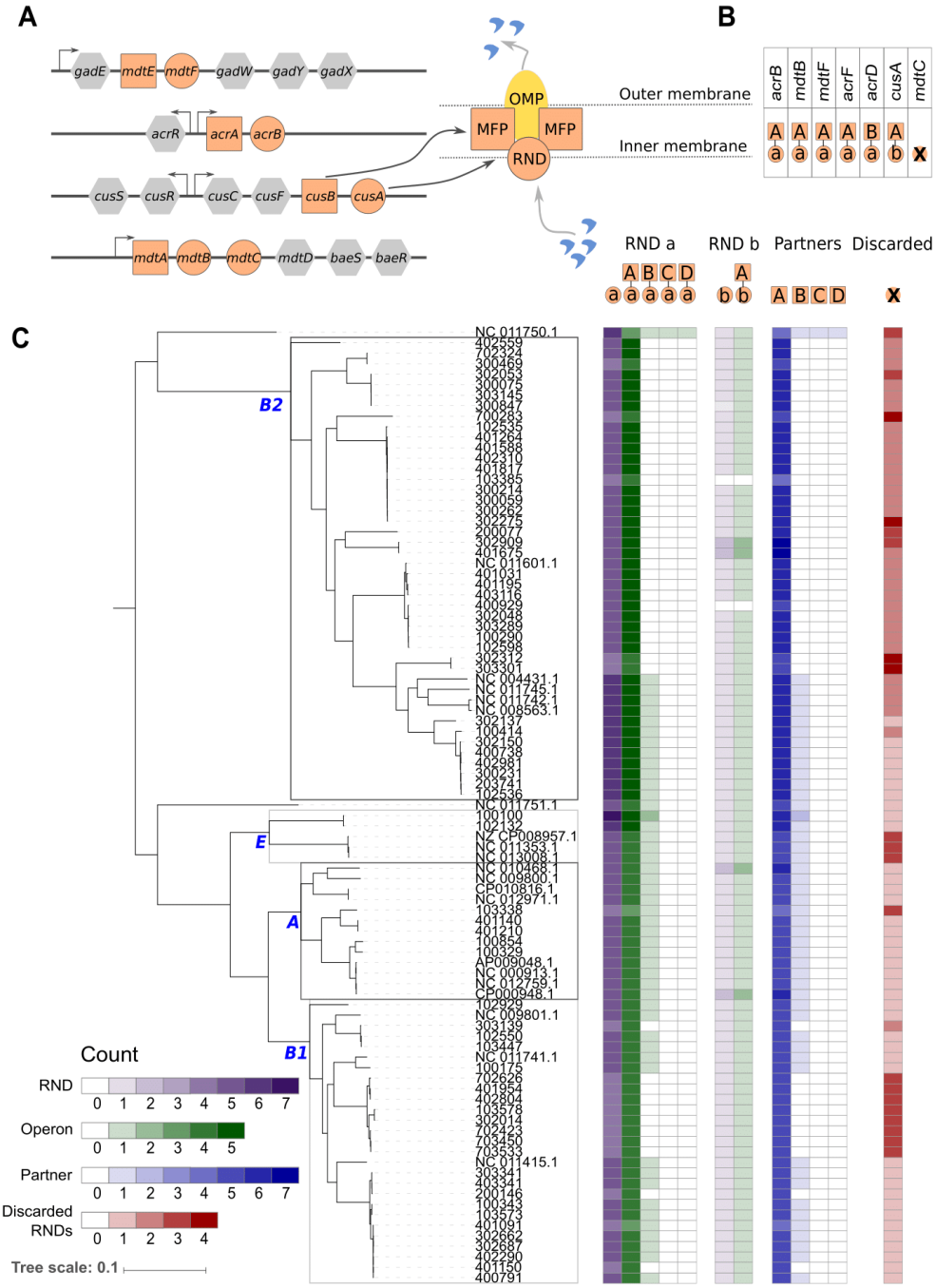

Figure 2.5: Identification of RND efflux pumps using SLING. A *Four example operon structures of RND efflux pumps present in* E. coli *K-12. All RND pump proteins share a single conserved HMM profile, represented by a single colour (ACR_tran;PF00873).* ***B*** *The corresponding annotation of RND efflux pumps in* E. coli *K-12 relative to the SLING output.* ***C*** *Annotation of RND efflux pumps in the* E. coli *collection. Darker squares represent presence of an RND pump protein or an operon in an isolate.*

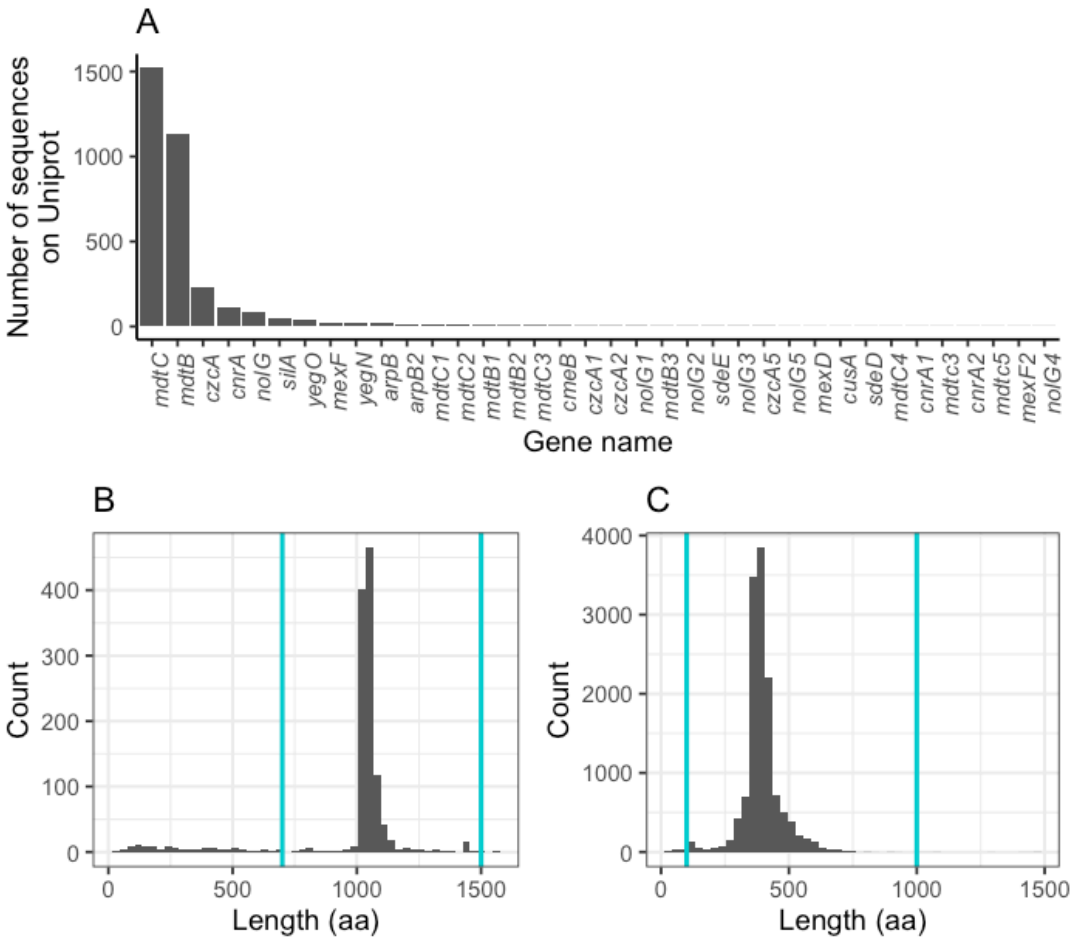

Figure 2.6: Defining the HMM collection and structural requirements for RND efflux pumps. A *Number of sequences retrieved from Uniprot using a name search of known RND efflux pumps genes.* ***B,C*** *Length distribution of RND efflux pump proteins (**B**) and MFPs (**C**) downloaded from Uniprot. Turquoise lines represent the cut-offs chosen as the length structural requirements for search using SLING.*

2.4.3.2 Benchmark on *E. coli* K-12

Seven RND efflux pumps are reported in the literature for *E. coli* K-12 strain W3110 (AP009048.1) [327]. Of these, SLING identified six RND pumps which fit the structure requirements applied in our analysis: *acrB*, *cusA*, *mdtB*, *acrF*, *acrD* and *mdtF* (Figure 2.5B). Since *mdtC* pump is found downstream to another RND pump, *mdtB*, (Figure 2.5A) this pump was discarded by SLING as the upstream gene was not in the correct length.

2.4.3.4 Application on EPEC collection

Five unique RND pump operons were identified in a SLING search on the collection of 90 EPEC and reference *E. coli* strains (Figure 2.5C). These operons consisted of two unique RND protein (hit) clusters (a and b) and four partner protein clusters (A-D).

The A partner protein is indeed an MFP and includes all the known MFPs found in *E. coli* K-12 (Figure 2.5B). It was highly prevalent and was observed in two different operons, with the two RND pump proteins (a and b). The “A-a” operon was ubiquitous, with at least four copies per strain. Reducing the identity threshold applied to group the proteins would have likely separated this operon into its corresponding operons in K-12. The “A-b” operon, on the other hand, was found in a single copy in most isolates. The “b” pump corresponded to the *cusA* RND pump in *E. coli* K-12, whereas the “a” pump represented all the other known RND pumps in *E. coli* K-12 (Figure 2.5B).

The B partner protein is a histidine kinase. This protein is identical in sequence to the *narQ* gene, found upstream to the *acrD* RND pump in *E. coli* K-12 [327]. This operon was missing in specific clades within the B1 and B2 phylogroups. These clades were correlated with the discarded hits, suggesting two events occurred that led to deviation from the expected operon structure in these clades.

Finally, the C and D partner proteins were only observed once and in a single isolate (ExPEC reference strain, *E. coli* IAI39). Both proteins were short with “C” partner protein 138 aa long and the “D” partner protein 310 aa long. BLAST results of protein “C” against the non-redundant protein sequence database suggest it is a histidine kinase similar to partner protein “B” (*narQ*). Protein “D”, on the other hand, is a truncated RND pump protein.

2.5 Discussion

SLING is an open source tool to examine the diversity of operons or gene arrays in bacterial datasets by using one of the conserved genes within the array to identify the linked genes which appear in a rule-defined proximity (Figure 2.7A). By examining the diversity of the neighbouring genes, we can elucidate incidences where there are deviations in the operon structure between isolates as well as deviations from what is expected to be the canonical operon structure of a specific system (Figure 2.7B). Examples of this were presented for the diversity of toxins as well as RND efflux pump proteins and their partner genes (i.e. antitoxins and MFPs) in a collection of *E. coli* isolates. While some genes presented a high diversity in their possible neighbours, others presented low diversity. Likewise, by examining the diversity of the neighbouring genes, SLING helped to further sub-categorise the gene combinations according to varying indications of these arrays being lost or gained.

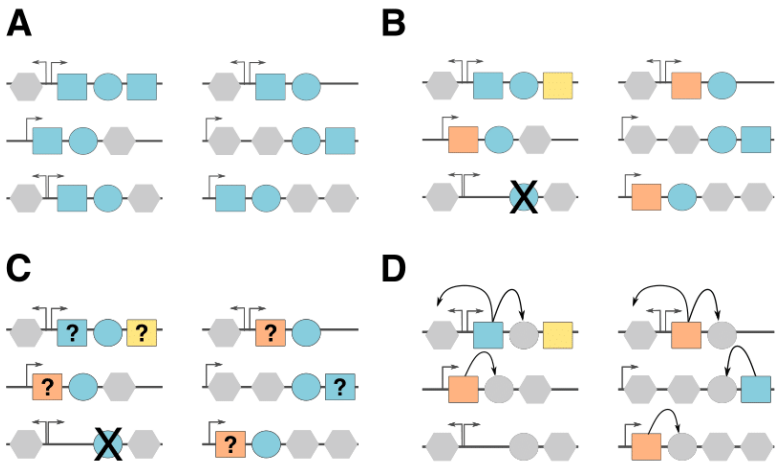

Figure 2.7: Utility of SLING. A *Search for gene pairs and triplets based on a single conserved gene (circle) and set of rules on the order and orientation of the neighbouring genes (squares)* ***B*** *Test the defined rules by examining the diversity of the neighbouring genes and identifying gene arrays which deviate from the expected structure* ***C*** *Directly identify new genes (squares)* ***D*** *Iteratively identify new genes by using the novel neighbour genes (squares) as the input HMM profiles.*

Two settings for TA systems and RND efflux pumps were described and these are built into the SLING interface for quick application using simple command line prompts, which are detailed on the tool's wiki page (https://github.com/ghoresh11/sling/wiki). Beyond these, SLING's advantage is in its flexibility; users can easily provide new profiles into its search, enabling identification of new and not well studied systems without relying on the developer to update the code or database. Thus, the utility of SLING is not limited to these operons and can be applied to other important operons or gene pairs such as CRISPR-cas systems, restriction-modification systems, secretion systems, and more. Users may construct HMM libraries and structural requirements in their area of expertise which can be shared with the community by uploading them to the public repository, enabling the extension of the built-in SLING use cases.

Additional advantage of SLING is that its protein search is based on an HMM profile search, rather than a sequence-based search, which allows SLING to capture more diverse members of a protein and not rely on a single sequence, likely taken from a lab strain which may not be representative in a collection of clinical or natural isolates. This advantage is also a limitation, it may be difficult to construct an HMM profile for an unknown gene when not many representative sequences are available. SLING also searches for the genes using a six-frame translation of the input genomes in addition to searching the CDSs predicted by annotation

tools. This allows the identification of short CDSs which may have otherwise been omitted by the annotation tools.

When searching for an unknown set of linked genes, SLING can also be used as a discovery tool. By applying default flexible structural requirements to find a partner gene, SLING can identify any set of genes which are linked to the primary gene. SLING can also be used to search for novel genes either directly, by looking at the partner genes identified (Figure 2.7C), or indirectly, but constructing HMM profiles of the newly identified partner genes and iteratively using these as the conserved gene (Figure 2.7D). These ideas are explored in Chapter 3 of this thesis, where SLING was used to examine the diversity of TA systems across a global collection of *K. pneumoniae* isolates.