T1TAdb: the database of Type I Toxin-Antitoxin systems

Nicolas J. Tourasse* and Fabien Darfeuille*

Univ. Bordeaux, CNRS, INSERM, ARNA, UMR 5320, U1212, F-33000 Bordeaux, France.

*corresponding authors. E-mail: nicolas.tourasse@inserm.fr; fabien.darfeuille@inserm.fr

Running head: Database of type I toxin-antitoxin systems

Keywords: toxin-antitoxin system, database, genome, antisense RNA, RNA structure, bioinformatics

Abstract

Type I toxin-antitoxin (T1TA) systems constitute a large class of genetic modules with antisense RNA (asRNA)-mediated regulation of gene expression. They are widespread in bacteria and consist of an mRNA coding for a toxic protein and a noncoding asRNA that acts as an antitoxin preventing the synthesis of the toxin by directly basepairing to its cognate mRNA. The co- and post-transcriptional regulation of T1TA systems is intimately linked to RNA sequence and structure, therefore it is essential to have an accurate annotation of the mRNA and asRNA molecules to understand this regulation. However, most T1TA systems have been identified by means of bioinformatic analyses solely based on the toxin protein sequences, and there is no central repository of information on their specific RNA features. Here we present the first database dedicated to type I TA systems, named T1TAdb. It is an open-access web database (https://d-lab.arna.cnrs.fr/t1tadb) with a collection of ~1,900 loci in ~500 bacterial strains in which a toxin-coding sequence has been previously identified. RNA molecules were annotated with a bioinformatic procedure based on key determinants of the mRNA structure and the genetic organization of the T1TA loci. Besides RNA and protein secondary structure predictions, T1TAdb also identifies promoter, ribosome-binding, and mRNA-asRNA interaction sites. It also includes tools for comparative analysis, such as sequence similarity search and computation of structural multiple alignments, which are annotated with covariation information. To our knowledge, T1TAdb represents the largest collection of features, sequences, and structural annotations on this class of genetic modules.

Introduction

Toxin-antitoxin (TA) systems are encoded within small genetic loci found in most of bacterial genomes including those of pathogens. They are usually composed of two adjacent genes: a stable toxin and a labile antitoxin that inhibits the toxin's action or expression and whose depletion rapidly leads to cell death or growth arrest (Harms et al. 2018). Six types of TA systems have been described so far depending on the nature and mode of action of the antitoxin (reviewed in (Goeders and Van

Melderen 2013)). While the toxin is always a protein, the antitoxin can be either a protein (types II, IV, V and VI) or an RNA (types I and III).

Type II TA systems are by far the most well studied class of TA systems. Extensive data for these systems are available in two databases, TADB 2.0 (Xie et al. 2018) and TASmania (Akarsu et al. 2019), which also include limited data for other types of TA systems. These databases also provide a tool (TAfinder and TASer, respectively) to scan and predict TA systems (Akarsu et al. 2019; Xie et al. 2018). Other TA-specific resources include BtoxDB (Barbosa et al. 2015), a database of TA protein structural data, and RASTA_Bacteria (Sevin and Barloy-Hubler 2007), a tool to scan for toxins and antitoxins in bacterial genomes (unfortunately, both are no longer maintained). Overall, none of these existing databases contains expanded data about type I TA (T1TA) systems.

A T1TA system consists of a relatively short mRNA (150 to 400 nucleotides long) coding for a small protein (20-60 amino acids in length) whose expression is toxic to the host cell and an antisense RNA (asRNA; 60-200 nt in length) that serves as a counteracting antitoxin to prevent the synthesis of its cognate toxin by directly basepairing to the mRNA. Numerous aspects of T1TA systems have been studied including RNA structure, toxin-antitoxin interaction, regulatory gene expression mechanisms (transcription, translation, degradation, processing), mechanism of action of the toxin, and function in cell physiology (Masachis and Darfeuille, 2018). While TA systems located on plasmids have been demonstrated to contribute to plasmid maintenance (via the mechanism of postsegregational killing, (Greenfield et al. 2000; Gerdes et al. 1986)), their role on the chromosome has only been addressed for a few TA loci (Brielle et al. 2016; Peltier et al. 2020).

A few T1TA systems have been experimentally characterized, but hundreds have been identified by bioinformatic analyses (Arnion et al. 2017; Fozo et al. 2010). Most of them are not yet annotated in genome records. In addition, the majority of loci have been identified solely on the basis of the toxin peptide sequence, while it is necessary to have an accurate annotation of the complete mRNA and asRNA molecules to understand the co- and post-transcriptional regulation of T1TA systems, which is determined by RNA sequence and structure (Masachis and Darfeuille 2018). Thus,

there is a deep need for a central repository of T1TA systems that would include RNA information. In this work, we have built a database, named T1TAdb, which gathers all described and predicted loci of T1TA systems and where the mRNA and asRNA coordinates are annotated. We describe below the content and main features of T1TAdb, along with the procedure used to identify mRNAs and asRNAs.

Results and Discussion

We have developed T1TAdb, the first database dedicated to T1TA systems. It can be accessed by users through a graphical web interface. A preliminary, development version was put on-line in January 2019 and has been regularly updated, corrected, and improved (https://d-lab.arna.cnrs.fr/t1tadb). The database provides sequence, secondary structure, and genomic information on T1TA loci. In its current, initial version, it is limited to bacterial genomes that were reported in the literature to carry T1TA systems. It contains 1891 loci belonging to 23 families from 216 bacterial species and 486 strains. Data on toxin mRNA, antitoxin asRNA, and toxin peptides were taken from the current literature. Only a small number of loci have been experimentally characterized and the bulk of the data in T1TAdb are mainly based on the results of genome-wide bioinformatic studies. This includes the AapA/IsoA family that has been extensively curated in about 100 genomes of *Helicobacter* and *Campylobacter* (Arnion et al. 2017) and the large number of loci found in the study of (Fozo et al. 2010). By exhaustive amino acid sequence homology searches using PSI-BLAST and TBLASTN run with customized parameters, Fozo and coworkers identified ~900 sequences of ORFs coding for homologs of known type I toxin peptides from 7 families in 95 bacterial species and 229 strains (Suppl. Table S5 in (Fozo et al. 2010)). In addition, through searches based on characteristics of T1TA loci (such as tandem repeats and hydrophobicity) they proposed more than 2,000 novel toxin ORFs in hundreds of genomes (Suppl. Table S8 therein). These analyses were essentially protein-based and did not investigate the mRNA genes. Based on thermodynamic local free energies, they could detect the position of asRNA genes of a few known families, but the

coordinates spanned only a 100-nt window and the precise start and end boundaries of the full-length molecules were not obtained (Suppl. Table S9 therein). Therefore, in our work, a major effort to build the T1TAdb database has been devoted to the annotation of the genomic locations of the toxin mRNA and antitoxin asRNA corresponding to each toxin ORF of known family reported by (Fozo et al. 2010), in order to reconstruct the complete loci.

Annotation of toxin mRNA and antitoxin asRNA genomic localization based on RNA structural features

The genomic regions defining the mRNAs and asRNAs were predicted using RNAMotif (Macke et al. 2001) and RNASurface (Soldatov et al. 2014), respectively. RNAMotif is used to identify regions that can adopt a predefined secondary structure, while RNASurface predicts regions that are structurally more stable than the rest of the genome. In our case, a complete TA locus was obtained when a pair of mRNA and asRNA could be predicted, with lengths, orientations, and relative positions matching the genetic organization of a given known TA family (Figure 1 and Suppl. Table S1; see Materials and Methods for details). As can be seen in Table 1, using this procedure, we were able to recover RNA regions of the expected family for 83 to 89% (depending on the family) of the toxin ORFs belonging to five of the seven known families detected by (Fozo et al. 2010) (Suppl. Table S5 therein). The success rate was however lower for the TxpA/RatA family (72%), and particularly low for the Fst/RNAII family (27%). This could be accounted by the high heterogeneity of the latter two families, for which a set of slightly different structural descriptors may be needed to accommodate specific subfamilies. In a few additional cases (less than 10% for each family), a locus could be recovered, but was assigned to a different family. For example, locus TA06047 from *Escherichia coli* classified in the Ldr/Rdl family (based on peptide features) was identified by our procedure (based on RNA and genetic organization features) as a member of the Hok/Sok family. This discrepancy could be due to a wrong family assignment by either bioinformatic method or peculiarities in the locus that make it look more similar to another family. The structural descriptors

that we used for RNAMotif searches were mainly based on key determinants of the toxin-encoding mRNA secondary structure, in particular Shine-Dalgarno (SD) and complementary anti-SD sequences, as well as terminator stem-loops in the 3' end, stem-loops in the 5' end, and regions involved in 5'-3' long-distance interactions. Because transcription and translation are coupled in bacteria, *cis*-encoded regulatory elements prevent premature translation of the toxin mRNA by sequestering the SD sequence, either co- or post-transcriptionally. Depending on the T1TA system, the anti-SD motif is located either at a short distance upstream of the SD sequence, thus creating a sequestering stem-loop, or near the end of the mRNA occluding the SD sequence via a long-distance base-pairing interaction between the 5' and 3' extremities of the mRNA (Masachis and Darfeuille 2018). The overall success of our results demonstrate that taking into account these elements, which are essential for the expression and regulation of T1TA systems, is critical for identifying type I toxin mRNAs.

Overview of T1TAdb features and tools

Information in T1TAdb can be searched by entering one or several keywords via the "Keyword Search" tab. A given keyword may be matched against any field in the database or against a specific field such as toxin or antitoxin name, species, strain, taxonomy, locus ID, or genome accession number. Submitting the search form will return a table listing all loci that match the search criteria. The "Select loci" button in the upper left corner of the table allows to further refine the selection of loci by replicon (chromosome or plasmid), TA family or taxonomy, and to select/deselect all loci. The table indicates the locus identifier (TA*nnnnn*), TA family, host strain, and reference publication of each locus. For loci that were predicted using structural descriptors of a TA family different from that reported in the literature, the predicted family is indicated, and groups of loci from the same genomic region matching multiple family characteristics appear under a specific identifier (TAG*nnnn*; "Overlapping Locus Group", see Materials and Methods for details). Clicking on a locus identifier leads to a "Locus Details" page that gives detailed information about the locus. The page is

divided into four panels (Figure 2). The "Locus" panel provides genomic information with an interactive full genome map (drawn with CGView; (Stothard and Wishart 2005)) of the host strain, showing the location of all TA loci harbored by the strain, and an embedded interactive genome browser (IGV; (Robinson et al. 2011); https://github.com/igvteam/igv.js/wiki) for viewing the genomic context around the locus. The "mRNA" and "sRNA" panels provide the sequence and secondary structure diagram of the toxin mRNA and antitoxin asRNA, respectively. The location of relevant motifs (start/stop codon, SD sequence, mRNA-asRNA interaction region) are annotated on the diagrams. The "Peptide" panel shows the sequence, secondary structure model and hydrophobicity plot of the toxin peptide. The "mRNA", "sRNA", and "Peptide" panels include a "BLAST Sequence" link for launching a sequence homology search (using BLAST+; (Camacho et al. 2009)) of the selected query against every mRNA, asRNA, ORF, peptide, or genome sequence in T1TAdb. The database also allows the user to input or upload one or several query sequences to perform a BLAST search via the "Sequence Search" tab.

The T1TAdb home page contains links to a full list of organisms ("view organism list") and TA families ("view family list") included in the database. Selecting a particular organism name or family name will return a table listing all loci that belong to that organism or family.

A notable implementation of T1TAdb is that, in addition to homology searches by BLAST, users can perform multiple sequence alignments (using MAFFT;(Katoh and Standley 2013)) to compare mRNA, asRNA, or peptide sequences. Alignments can be launched via the "Align selected sequences" button on top of the table that is returned after selecting loci by organism or family, or following a keyword search. An embedded interactive alignment viewer (MSAViewer; (Yachdav et al. 2016)) is provided to browse and download the alignments. For RNA sequences, structural information is taken into account in MAFFT to produce a structural alignment (Katoh and Toh 2008), and a consensus 2D structure is predicted (using RNAalifold; (Bernhart et al. 2008)) from the alignment. Furthermore, a covariation analysis is performed (using R-chie; (Lai et al. 2012)) to

annotate the alignment with covariation information ((Tourasse and Darfeuille 2020); dedicated links are provided to download the consensus and covariation data).

In addition to alignment results, the other data in T1TAdb are also available for download. Once a set of loci has been selected, the corresponding data can be downloaded using the "Download selected loci" menu at the top of the loci table. Users can obtain a spreadsheet file containing the detailed information about the loci (including host organism, genomic coordinates and sequences of the RNA, ORF, peptide, and promoter features) or sequence files in FASTA format. In the "Locus Details" page for a given locus the structure diagrams can be saved in various formats (SVG, PDF, PNG, or GIF).

The list of publications from which sequences, structures, and other information about the T1TA systems have been obtained is given in the "References" page. Links to other resources on TA systems, as well as the various software and tools used to build T1TAdb, are provided in the "Links" page.

Example of insight gained from the use of T1TAdb

We present below an example illustrating the usefulness and added value provided by T1TAdb. A structural alignment of the Sok antitoxin asRNAs from the Hok/Sok family, which is present in various Gram-negative bacteria, revealed two subgroups: one subgroup had a 2D structure conforming to that reported by (Franch et al. 1997) (Figure 3A), while the other carries an additional stem-loop element at the 5' end (Figure 3B). This extra element is well supported by covarying positions that reveal compensatory substitutions in a number of sequences, and also by the presence of a -10 promoter box motif (TANNNT, where N means any nucleotide) at an appropriate location upstream of the stem-loop. Although these observations should be verified experimentally, it shows the potential of T1TAdbto reveal new features in RNA-mediated regulation of T1TA systems that could not be identified during the study of single T1TA loci.

Conclusions and perspectives

In this work, we have built the first web database totally dedicated to T1TA systems, named T1TAdb. T1TA systems are widespread in bacteria, including pathogenic species, and are studied for numerous aspects including RNA and protein structure and function, regulation of gene expression and cell physiology. In addition to providing access to the collection of loci that have been reported in the literature, the database brings an added value by the annotation of toxin mRNAs and antitoxin asRNAs of loci that were predicted solely by analysis of the toxin peptide sequence. This information, together with a wealth of sequence, structure, and genomic data provided on T1TA systems, along with the other databases and tools dedicated to TA systems (such as TADB and TASmania), will certainly serve the scientific community to gain deeper insights of the distribution, evolution, structure, and function of TA systems. This could reveal to be particularly important for the TA systems located in bacterial chromosomes, whose function remains largely unknown.

In future developments of T1TAdb, we plan to expand the collection of T1TA systems by predicting TA loci in all bacterial genomes available. In addition, the procedure described in this work for the annotation of mRNAs and asRNAs of T1TA systems, which is based on structural features, can be turned into an automatic *de novo* prediction tool, most likely with additional steps to control the false-positive rate. The RNA data in T1TAdb may also be used to generate alignments and covariance models in order to search genome databases for conserved RNAs, as exemplified in the Rfam database (Kalvari et al. 2021). We also anticipate to incorporate transcriptome (RNA-Seq) data into T1TAdb. As the number of sequenced bacterial transcriptomes is increasing, such information would be valuable to check whether TA loci are expressed and to verify the genomic coordinates of the predicted RNAs. Further improvements to T1TAdb will also include tools for simultaneously visualizing and comparing the genomic context (gene neighborhoods) across multiple strains or for the reconstruction of phylogenetic trees, and will allow users to submit data for inclusion in T1TAdb.

Availability

T1TAdb is freely available on-line at https://d-lab.arna.cnrs.fr/t1tadb.

Materials and Methods

Data collection

Sequence, secondary structure, and genomic localization of T1TA loci (including mRNA, asRNA, ORF, and promoter -10 boxes) from the various families were taken from the publications where they were initially discovered and/or characterized (Arnion et al. 2017; Durand et al. 2012; Fozo et al. 2010; Weaver et al. 2009; Maikova et al. 2018; Folli et al. 2017; Kristiansen et al. 2016; Wen and Fozo 2014; Pinel-Marie et al. 2014; Wen et al. 2014; Jahn and Brantl 2013; Weaver 2012; Fozo 2012; Sayed et al. 2012; Han et al. 2010; Sharma et al. 2010; Darfeuille et al. 2007; Pichon and Felden 2005; Kawano et al. 2002; Pedersen and Gerdes 1999; Franch et al. 1997; Masachis and Darfeuille 2018; Meißner et al. 2016; Peltier et al. 2020; Germain-Amiot et al. 2019). In most cases, reports were on one or few copies of a given TA system, identified or studied in one or a few bacterial strains. Information on TA systems belonging to the AapA/IsoA family identified by bioinformatic analyses and manual curation in ~100 genomes of *Helicobacter* and *Campylobacter* were taken from (Arnion et al. 2017). (Sharma et al. 2010) reported members of the AapB/IsoB, AapC/IsoC, and AapD/IsoD families in six *Helicobacter* strains. For the sake of consistency and completeness, we located the homologues of these loci in all other *Helicobacter* genomes screened by (Arnion et al. 2017) through sequence homology searches (using BLAST+ 2.2.31; (Camacho et al. 2009)) and multiple sequence alignment. (Kristiansen et al. 2016) identified mRNAs and ORFs belonging to the DinQ/AgrB family in 15 Gram-negative bacteria and we devised a procedure to predict the corresponding asRNAs (see below). The majority of the data in T1TAdb are based on the genome-wide bioinformatic searches by (Fozo et al. 2010) who identified sequences of ORFs coding for type I toxin peptides of known and novel families in hundreds of bacterial strains. For toxins belonging to known TA families, we performed computational analyses to locate the coordinates of the toxin

mRNA and antitoxin asRNA in order to identify the complete TA locus corresponding to each reported toxin ORF.

Prediction of toxin mRNA and antitoxin asRNA genomic localization

Identification of mRNAs and asRNAs was based on the characteristics of known examples. The prediction pipeline is summarized in Fig. 1. Genomic locations of toxin mRNAs were determined using RNAMotif 3.1.1 (Macke et al. 2001). RNAMotif takes as input a descriptor file that contains parameters describing the various elements that make up the secondary structure of a particular RNA (helices, loops, etc. and their respective lengths). Using this descriptor RNAMotif scans genome sequences to find regions that can adopt (i.e., can be folded into) the specified structure. RNAMotif descriptors for mRNAs belonging to 16 known T1TA families were written based on sequences and secondary structures reported in the literature cited above. We did not include every structural element in the descriptors, but rather focused on key determinants such as SD and anti-SD sequences (that sequester the SD motif), location of the ORF within the mRNA, stem-loops in the 5' end or terminator stem-loops in the 3' end, and regions implicated in 5'-3' long-distance interactions (textual summary of RNAMotif parameters given in Suppl. Table S1). The remaining parts of the structure were specified as undefined regions. To avoid descriptors being too specific to a given mRNA instance, lengths and distances between the various elements were usually not set to a defined value but to a min./max. value or a relatively broad range of values. All genomes surveyed by (Fozo et al. 2010) (genome sequences downloaded from NCBI RefSeq; (Haft et al. 2018)) were scanned by RNAMotif with descriptors of all 16 TA families. Hits that spanned the coordinates of the ORFs reported by Fozo *et al.* on the same DNA strand were extracted (by means of the "intersectbed" utility from the BEDTools 2.24.0 package (Quinlan and Hall 2010)). Among those, hits that contained ORFs that were in-frame with the ORFs of Fozo *et al.* (ORFs may not always be identical and could differ slightly in length in some cases depending on the start codon used) and whose ORF and total mRNA

length matched the typical range of ORF and mRNA length for a given family were retrieved, whether or not the family matched that reported by Fozo and coworkers.

Antitoxin asRNAs were localized using RNASurface 1.1 (Soldatov et al. 2014), which predicts sequence regions that are structurally more stable than the rest of the genome by local folding of segments up to a predefined size. Genome sequences were scanned with RNASurface (run with options "--winmin 50 --winmax 300 -z -2 -d 500") to identify structured segments of length 50-300 nt that have a z-score lower than -2 (corresponding to a false-positive rate of 5%, (Soldatov et al. 2014)). For a given TA family, RNASurface segments whose length corresponded to the typical asRNA length for that family (+/- 20%) were retrieved.

Putative promoter -10 boxes were predicted by searching (by means of the PERL module Regexp::Exhaustive) for the presence of the sequence motif TANNNT (where N means any nucleotide) in a window covering the region -20 to -6 upstream of the identified mRNAs and asRNAs. In case where multiple motifs were present the 3'-most (i.e., the closest to the RNA start) was selected.

In order to find the pair of mRNA and asRNA corresponding to the same TA locus, RNAMotif and RNASurface results were intersected (using "intersectbed" from BEDTools) based on the defined genetic organization of the various T1TA families. In the ShoB/OhsC, TisB/IstR1, Zor/Orz, and DinQ/AgrB families, found in Gram-negative bacteria, the asRNA is located 5' to the mRNA and does not overlap it, whereas in the Fst/RNAII, TxpA/RatA, YonT/as-YonT, SprA1/SprA1as, SprG/SprF, and BsrE-G-H/as-BsrE-G-H families, all found in Gram-positive bacteria, the asRNA is located on the 3' side of the mRNA and partially overlaps it (overlaps also the ORF except in the Fst/RNAII family); the asRNA is fully overlapped by the mRNA in the AapA/IsoA, Ibs/Sib, Hok/Sok, and Ldr/Rdl families from Gram-negative organisms, but overlaps the ORF only in the AapA/IsoA and Ibs/Sib families (spans the entire ORF in Ibs/Sib), whereas it is located 5' of the ORF in the Hok/Sok and Ldr/Rdl families ((Wen and Fozo 2014; Arnion et al. 2017; Pinel-Marie et al. 2014; Durand et al. 2012; Sayed et al. 2012); Suppl. Table S1). For families where the two RNAs do

not overlap, a max. distance of 300 nt between the RNAs was allowed (except for the DinQ/AgrB family, see below), and for families where the two RNAs do overlap, min. and max. limits were set for the length of the overlap region based on known examples (Suppl. Table S1). In T1TA systems the mRNA and asRNA share a complementary region of interaction. In families with overlapping RNAs this region generally corresponds to the overlap region, but for the Fst/RNAII and SprA1/SprA1as families interaction occurs outside the overlap sequence (Weaver et al. 2009; Sayed et al. 2012). The software IntaRNA 3.1.0.2 (Mann et al. 2017) was used to find mRNA-asRNA pairs that share a complementary region at the expected location for these two families and to identify regions of interaction for families with non-overlapping RNAs (the interaction region was set to be at least 15 nt long). A specific processing was carried out for the DinQ/AgrB family because in some bacterial strains the locus includes an AgrA ncRNA gene that is homologous to the AgrB asRNA and that can be located in-between DinQ and AgrB. In *Escherichia coli* K-12 substr. MG1655 the sequence of the interaction region in AgrA contains a number of mismatches and it has been shown that AgrA does not bind the DinQ mRNA and that only AgrB, which shares a region almost fully complementary to DinQ, acts as an antitoxin (Kristiansen et al. 2016). Therefore, to accommodate for the possible presence of the AgrA and AgrB paralogs, the max. distance between the mRNA and asRNA was extended from 300 to 500 nt. In cases where two asRNAs were predicted in this region by RNASurface and for both of them a possible sequence of interaction (≥ 15 nt) with the mRNA was identified by IntaRNA, we selected the one for which the interaction was predicted to be the most stable (i.e., had the lowest free energy as computed by IntaRNA). The same procedure was followed to predict the asRNAs corresponding to the DinQ-like loci reported in (Kristiansen et al. 2016) where only mRNAs and ORFs were annotated.

Following the characteristics described above, all pairs of mRNAs and asRNAs that were in the correct orientation and distance and that shared a complementary interaction region were identified for each family. This set of pairs was then intersected with the set of mRNAs that were found to encode ORFs corresponding to those reported by (Fozo et al. 2010), to obtain the pairs that

include a known ORF (Fig. 1). For a given locus, there were usually several mRNAs and asRNAs matching these criteria because we used a relatively broad range of values for the parameters describing the length, structure, and organization of the RNAs. Due to the flexibility in the RNAMotif structural descriptors multiple overlapping mRNAs spanning the same ORF were always found, differing by their lengths and a few bases in their start/end coordinates. Among those, the one for which a promoter -10 box could be predicted was selected as the mRNA for the given locus. If a promoter was predicted for multiple mRNAs, or if no promoter was found for any of the mRNAs, then RNA sequences were folded using MFOLD 3.6 (Zuker 1989, 2003) and the minimum free energy (MFE) of the most stable structure of each RNA was normalized by sequence length (adjusted MFE; AMFE) to compare the stability among RNAs. The mRNA that had the smallest AMFE was taken as the mRNA for the locus. The use of AMFE was warranted by the fact that mRNA lengths of the different T1TA families are in the range 200-400 nt where AMFE is almost length-independent (Trotta 2014). For some loci, there were also several possible asRNAs matching the TA family characteristics as RNASurface often predicts multiple RNA segments of different lengths overlapping the same genomic region. Among those, the one for which a promoter -10 box was predicted was selected as the asRNA for the locus. If a promoter was predicted for multiple asRNAs, or if no promoter was found for any of the asRNAs, then the one with the lowest RNASurface z-score was chosen as the putative asRNA for the given locus. Z-score was used here as normalized folding stability measure because asRNA lengths are in the range 60-200 nt where AMFE is length-dependent and thus cannot be reliably used to compare stabilities among RNAs (Trotta 2014).

To verify that the correct asRNA was associated with a given mRNA, an alignment of the antitoxin sequences was made for each family to check the homology and completeness of the sequences. The alignment included only the RNAs whose family corresponded to the family of the ORF in (Fozo et al. 2010), in which case one should expect these sequences to be homologous to each other. Incomplete sequences or sequences whose coordinates were shifted relative to the others

were manually corrected. No alignment was done for the mRNAs because they are constrained by the location of the ORFs and thus are normally homologous.

If no suitable mRNA-asRNA pair spanning an ORF identified by (Fozo et al. 2010) could be found, that particular ORF was not included in T1TAdb. However, in cases where no mRNA-asRNA pair corresponding to the same family as that assigned to the ORF in (Fozo et al. 2010) could be identified but the ORF was included in an RNA pair of a different family, the locus was retained and a "predicted_family" flag was set to indicate that it matched an alternative family. This flag was also set when ORFs from unknown or novel families were part of loci corresponding to known families. If RNA pairs of multiple TA families spanned the same ORF, only the one matching the family of the ORF was retained. If there were no such pair, then alternative loci were filtered according to the type of organism in which the ORF was encoded, i.e., for a Gram-negative bacterium only the RNA pairs from families found in Gram-negative bacteria were retained, and similarly for Gram-positive organisms. If there were no such pairs, then all alternative mRNA-asRNA pairs covering the ORF were retained. The multiple loci spanning a given ORF were organized into an "Overlapping Locus Group", which represents a group of loci that have been identified using characteristics of different TA families and that overlap the same genomic region, but we could not determine which one is the real locus in this region.

Database implementation

T1TAdb is implemented as a relational database in PostgreSQL 10.13 (https://www.postgresql.org/). The database schema was designed following the five rules of normalization (http://www.barrywise.com/2008/01/database-normalization-and-design-techniques/) to avoid data redundancy and inconsistent dependency among tables. The graphical web interface was developed using the PERL Catalyst framework (http://www.catalystframework.org/) to control and manage connections and SQL requests to the database. Graphical design of the web pages was done in dynamic and responsive HTML, JavaScript, and Cascading Style Sheets (CSS)

(https://www.w3schools.com/), along with the PERL Template Toolkit 2.26 templating system (http://www.template-toolkit.org/). The T1TAdb website is run via the Apache HTTP 2.4.6 server (https://httpd.apache.org/) under the Linux CentOS 7.8 operating system. Secondary structures of RNA were predicted using MFOLD 3.6 (Zuker 2003, 1989) and annotated diagrams highlighting the location of specific motifs (start/stop codon, SD sequence, interaction region)were generated with VARNA 3.93 (Darty et al. 2009). Secondary structures of toxin peptides were predicted using PSIPRED 4.02 (McGuffin et al. 2000) run with PSI-BLAST 2.2.26 against the UniRef90 protein sequence database (https://www.uniprot.org/help/uniref) and drawn with POLYVIEW-2D (Porollo et al. 2004). Hydrophobicity plots were computed with ProtScale (Gasteiger et al. 2005). Interactive genomic maps in SVG format showing the localizations of TA loci were drawn using CGView (Stothard and Wishart 2005). The embeddable JavaScript/CSS version of the IGV browser ((Robinson et al. 2011); https://github.com/igvteam/igv.js/wiki) was used to provide an interactive visualization of the genomic context flanking TA loci, and SVG images of the genomic context were generated by means of the Gviz package (https://bioconductor.org/packages/release/bioc/html/Gviz.html) in R 3.5.3 ((R Development Core Team 2019); https://www.r-project.org/). Sequence similarity searches in T1TAdb are done with BLAST+ 2.2.31 (Camacho et al. 2009) and multiple sequence alignments are computed by MAFFT 7.407 (Katoh and Standley 2013). For peptide sequences MAFFT is run with the method “mafft-linsi” and the option “--localpair”, whereas for RNA sequences MAFFT is run with the method “mafft-xinsi” and the option “--scarnapair” to incorporate structure information and produce a structural alignment (Katoh and Toh 2008). Alignments are visualized using the embeddable MSAViewer (Yachdav et al. 2016), which is part of the BioJS JavaScript tools (Yachdav et al. 2015). In addition, for RNA alignments, a consensus structure is predicted by means of RNAalifold from the ViennaRNA 2.1.9 package (Lorenz et al. 2011; Bernhart et al. 2008) and a covariation analysis is conducted using R-chie (Lai et al. 2012). Other bioinformatic software such as EMBOSS 6.6.0 (Rice et al. 2000), BioPERL (Stajich et al. 2002), and FASTX-Toolkit 0.0.14

(http://hannonlab.cshl.edu/fastx_toolkit/) were also employed to generate and/or process the data included in T1TAdb.

Acknowledgements

This work was supported by Institut National de la Santé et de la Recherche Médicale [INSERM U1212], Centre National de la Recherche Scientifique [CNRS UMR5320], Bordeaux University, and Agence Nationale de la Recherche [ANR-12-BSV5-0025-Bactox1, ANR-12-BSV6-0007-asSUPYCO]. The web server hosting T1TAdb is provided by the ODS Web Hosting service of CNRS. We thank Dr. Stéphane Thore and Dr. Sébastien Fribourg for providing additional computing power. Finally, we thank all present and past members of the Darfeuille laboratory for fruitful discussions on this project and Anthony Bugaut, Isabelle Iost, Anaïs Le Rhun, Simon Bonabal, and Olga Soutourina for critical comments on the manuscript.

References

Akarsu H, Bordes P, Mansour M, Bigot DJ, Genevaux P, Falquet L. 2019. TASmania: A bacterial toxin-antitoxin systems database. *PLoS Comput Biol* **15**: e1006946.

Arnion H, Korkut DN, Gelo SM, Chabas S, Reignier J, Iost I, Darfeuille F. 2017. Mechanistic insights into type I toxin antitoxin systems in Helicobacter pylori: The importance of mRNA folding in controlling toxin expression. *Nucleic Acids Res* **45**: 4782–4795.

Barbosa LCB, Garrido SS, Marchetto R. 2015. BtoxDB: A comprehensive database of protein structural data on toxin-antitoxin systems. *Comput Biol Med* **58**: 146–153. http://dx.doi.org/10.1016/j.compbiomed.2015.01.010.

Bernhart SH, Hofacker IL, Will S, Gruber AR, Stadler PF. 2008. RNAalifold: Improved consensus structure prediction for RNA alignments. *BMC Bioinformatics* **9**: 474.

Brielle R, Pinel-Marie M-L, Felden B. 2016. Linking bacterial type I toxins with their actions. *Curr Opin Microbiol* **30**: 114–121. https://pubmed.ncbi.nlm.nih.gov/26874964/ (Accessed March

23, 2021).

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009. BLAST+: Architecture and applications. *BMC Bioinformatics* **10**: 1–9.

Darfeuille F, Unoson C, Vogel J, Wagner EGH. 2007. An Antisense RNA Inhibits Translation by Competing with Standby Ribosomes. *Mol Cell* **26**: 381–392.

Darty K, Denise A, Ponty Y. 2009. VARNA: Interactive drawing and editing of the RNA secondary structure. *Bioinformatics* **25**: 1974–1975.

Durand S, Jahn N, Condon C, Brantl S. 2012. Type I toxin-antitoxin systems in Bacillus subtilis. *RNA Biol* **9**: 1491–1497.

Folli C, Levante A, Percudani R, Amidani D, Bottazzi S, Ferrari A, Rivetti C, Neviani E, Lazzi C. 2017. Toward the identification of a type i toxin-antitoxin system in the plasmid DNA of dairy Lactobacillus rhamnosus. *Sci Rep* **7**: 1–13.

Fozo EM. 2012. New type I toxin-antitoxin families from “wild” and laboratory strains of E. coli: Ibs-Sib, ShoB-OhsC and Zor-Orz. *RNA Biol* **9**: 1504–1512.

Fozo EM, Makarova KS, Shabalina SA, Yutin N, Koonin E V., Storz G. 2010. Abundance of type I toxin-antitoxin systems in bacteria: Searches for new candidates and discovery of novel families. *Nucleic Acids Res* **38**: 3743–3759.

Franch T, Gultyaev AP, Gerdes K. 1997. Programmed cell death by hok/sok of plasmid R1: processing at the hok mRNA 3’-end triggers structural rearrangements that allow translation and antisense RNA binding. *J Mol Biol* **273**: 38–51. http://www.ncbi.nlm.nih.gov/pubmed/9367744.

Gasteiger E, Hoogland C, Gattiker A, Duvaud S, Wilkins MR, Appel RD, Bairoch A. 2005. The Proteomics Protocols Handbook. 571–608. http://link.springer.com/10.1385/1592598900.

Gerdes K, Rasmussen PB, Molin S. 1986. Unique type of plasmid maintenance function: postsegregational killing of plasmid-free cells. *Proc Natl Acad Sci* **83**: 3116–3120. http://www.pnas.org/cgi/doi/10.1073/pnas.83.10.3116.

Germain-Amiot N, Augagneur Y, Camberlein E, Nicolas I, Lecureur V, Rouillon A, Felden B. 2019. A novel Staphylococcus aureus cis-trans type I toxin-antitoxin module with dual effects on bacteria and host cells. *Nucleic Acids Res* **47**: 1759–1773.

Goeders N, Van Melderen L. 2013. Toxin-antitoxin systems as multilevel interaction systems. *Toxins (Basel)* **6**: 304–324.

Greenfield TJ, Ehli E, Kirshenmann T, Franch T, Gerdes K, Weaver KE. 2000. The antisense RNA of the par locus of pAD1 regulates the expression of a 33-amino-acid toxic peptide by an unusual mechanism. *Mol Microbiol* **37**: 652–660.

Haft DH, DiCuccio M, Badretdin A, Brover V, Chetvernin V, O'Neill K, Li W, Chitsaz F, Derbyshire MK, Gonzales NR, et al. 2018. RefSeq: An update on prokaryotic genome annotation and curation. *Nucleic Acids Res* **46**: D851–D860.

Han K, Kim KS, Bak G, Park H, Lee Y. 2010. Recognition and discrimination of target mRNAs by Sib RNAs, a cis-encoded sRNA family. *Nucleic Acids Res* **38**: 5851–5866.

Harms A, Brodersen DE, Mitarai N, Gerdes K. 2018. Toxins, Targets, and Triggers: An Overview of Toxin-Antitoxin Biology. *Mol Cell* **70**: 768–784. http://www.ncbi.nlm.nih.gov/pubmed/29398446.

Jahn N, Brantl S. 2013. One antitoxin-two functions: SR4 controls toxin mRNA decay and translation. *Nucleic Acids Res* **41**: 9870–9880.

Kalvari I, Nawrocki EP, Ontiveros-Palacios N, Argasinska J, Lamkiewicz K, Marz M, Griffiths-Jones S, Toffano-Nioche C, Gautheret D, Weinberg Z, et al. 2021. Rfam 14: expanded coverage of metagenomic, viral and microRNA families. *Nucleic Acids Res* **49**: D192–D200.

Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Mol Biol Evol* **30**: 772–780.

Katoh K, Toh H. 2008. Improved accuracy of multiple ncRNA alignment by incorporating structural information into a MAFFT-based framework. *BMC Bioinformatics* **9**: 212.

Kawano M, Oshima T, Kasai H, Mori H. 2002. Molecular characterization of long direct repeat

(LDR) sequences expressing a stable mRNA encoding for a 35-amino-acid cell-killing peptide and a cis-encoded small antisense RNA in Escherichia coli. *Mol Microbiol* **45**: 333–349.

Kristiansen KI, Weel-Sneve R, Booth JA, Bjørås M. 2016. Mutually exclusive RNA secondary structures regulate translation initiation of DinQ in Escherichia coli. *RNA* **22**: 1739–1749.

Lai D, Proctor JR, Zhu JYA, Meyer IM. 2012. R-CHIE: A web server and R package for visualizing RNA secondary structures. *Nucleic Acids Res* **40**: e95.

Lorenz R, Bernhart SH, Höner zu Siederdissen C, Tafer H, Flamm C, Stadler PF, Hofacker IL. 2011. ViennaRNA Package 2.0. *Algorithms Mol Biol* **6**: 26.

Macke TJ, Ecker DJ, Gutell RR, Gautheret D, Case DA, Sampath R. 2001. RNAMotif, an RNA secondary structure definition and search algorithm. *Nucleic Acids Res* **29**: 4724–4735. https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/29.22.4724.

Maikova A, Peltier J, Boudry P, Hajnsdorf E, Kint N, Monot M, Poquet I, Martin-Verstraete I, Dupuy B, Soutourina O. 2018. Discovery of new type I toxin–antitoxin systems adjacent to CRISPR arrays in Clostridium difficile. *Nucleic Acids Res* **46**: 4733–4751. https://academic.oup.com/nar/advance-article/doi/10.1093/nar/gky124/4909981.

Mann M, Wright PR, Backofen R. 2017. IntaRNA 2.0: Enhanced and customizable prediction of RNA-RNA interactions. *Nucleic Acids Res* **45**: W435–W439.

Masachis S, Darfeuille F. 2018. Type I Toxin-Antitoxin Systems: Regulating Toxin Expression via Shine-Dalgarno Sequence Sequestration and Small RNA Binding. *Regul with RNA Bact Archaea* 173–190. http://www.asmscience.org/content/book/10.1128/9781683670247.chap11.

McGuffin LJ, Bryson K, Jones DT. 2000. The PSIPRED protein structure prediction server. *Bioinformatics* **16**: 404–405.

Meißner C, Jahn N, Brantl S. 2016. In Vitro Characterization of the Type I Toxin-Antitoxin System bsrE/SR5 from Bacillus subtilis. *J Biol Chem* **291**: 560–71. http://www.ncbi.nlm.nih.gov/pubmed/26565032.

Pedersen K, Gerdes K. 1999. Multiple hok genes on the chromosome of Escherichia coli. *Mol*

Microbiol **32**: 1090–1102.

Peltier J, Hamiot A, Garneau JR, Boudry P, Maikova A, Hajnsdorf E, Fortier LC, Dupuy B, Soutourina O. 2020. Type I toxin-antitoxin systems contribute to mobile genetic elements maintenance in Clostridioides difficile and can be used as a counter-selectable marker for chromosomal manipulation. *Commun Biol* **3**: 718. https://doi.org/10.1038/s42003-020-01448-5.

Pichon C, Felden B. 2005. Small RNA genes expressed from Staphylococcus aureus genomic and pathogenicity islands with specific expression among pathogenic strains. *Proc Natl Acad Sci U S A* **102**: 14249–14254. http://dx.doi.org/10.1073/pnas.0503838102.

Pinel-Marie M-L, Brielle R, Felden B. 2014. Dual Toxic-Peptide-Coding Staphylococcus aureus RNA under Antisense Regulation Targets Host Cells and Bacterial Rivals Unequally. *Cell Rep* **7**: 424–435.

Porollo AA, Adamczak R, Meller J. 2004. POLYVIEW: A flexible visualization tool for structural and functional annotations of proteins. *Bioinformatics* **20**: 2460–2462.

Quinlan AR, Hall IM. 2010. BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**: 841–842.

R Development Core Team. 2019. R: A language and environment for statistical computing. *R Found Stat Comput*. https://www.r-project.org/.

Rice P, Longden I, Bleasby A. 2000. EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet* **16**: 276–7. http://www.ncbi.nlm.nih.gov/pubmed/10827456.

Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP. 2011. Integrative genomics viewer. *Nat Biotechnol* **29**: 24–6. http://www.ncbi.nlm.nih.gov/pubmed/21221095.

Sayed N, Jousselin A, Felden B. 2012. A cis-antisense RNA acts in trans in Staphylococcus aureus to control translation of a human cytolytic peptide. *Nat Struct Mol Biol* **19**: 105–113. http://dx.doi.org/10.1038/nsmb.2193.

Sevin EW, Barloy-Hubler F. 2007. RASTA-Bacteria: A web-based tool for identifying toxin-antitoxin loci in prokaryotes. *Genome Biol* **8**.

Sharma CM, Hoffmann S, Darfeuille F, Reignier J, Findeiß S, Sittka A, Chabas S, Reiche K, Hackermüller J, Reinhardt R, et al. 2010. The primary transcriptome of the major human pathogen Helicobacter pylori. *Nature* **464**: 250–255.

Soldatov RA, Vinogradova S V., Mironov AA. 2014. RNASurface: Fast and accurate detection of locally optimal potentially structured RNA segments. *Bioinformatics* **30**: 457–463.

Stajich JE, Block D, Boulez K, Brenner SE, Chervitz SA, Dagdigian C, Fuellen G, Gilbert JGR, Korf I, Lapp H, et al. 2002. The Bioperl toolkit: Perl modules for the life sciences. *Genome Res* **12**: 1611–1618.

Stothard P, Wishart DS. 2005. Circular genome visualization and exploration using CGView. *Bioinformatics* **21**: 537–539.

Tourasse N, Darfeuille F. 2020. Structural Alignment and Covariation Analysis of RNA Sequences. *Bio-Protocol* **10**: e3511.

Trotta E. 2014. On the normalization of the minimum free energy of RNAs by sequence length. *PLoS One* **9**.

Weaver KE. 2012. The par toxin-antitoxin system from Enterococcus faecalis plasmid pAD1 and its chromosomal homologs. *RNA Biol* **9**: 1498–1503.

Weaver KE, Reddy SG, Brinkman CL, Patel S, Bayles KW, Endres JL. 2009. Identification and characterization of a family of toxin-antitoxin systems related to the Enterococcus faecalis plasmid pAD1 par addiction module. *Microbiology* **155**: 2930–2940.

Wen J, Fozo EM. 2014. sRNA antitoxins: More than one way to repress a toxin. *Toxins (Basel)* **6**: 2310–2335.

Wen J, Won D, Fozo EM. 2014. The ZorO-OrzO type I toxin-Antitoxin locus: Repression by the OrzO antitoxin. *Nucleic Acids Res* **42**: 1930–1946.

Xie Y, Wei Y, Shen Y, Li X, Zhou H, Tai C, Deng Z, Ou HY. 2018. TADB 2.0: An updated

database of bacterial type II toxin-antitoxin loci. *Nucleic Acids Res* **46**: D749–D753.

Yachdav G, Goldberg T, Wilzbach S, Dao D, Shih I, Choudhary S, Crouch S, Franz M, García A, García LJ, et al. 2015. Anatomy of BioJS, an open source community for the life sciences. *Elife* **4**: 1–7.

Yachdav G, Wilzbach S, Rauscher B, Sheridan R, Sillitoe I, Procter J, Lewis SE, Rost B, Goldberg T. 2016. MSAViewer: Interactive JavaScript visualization of multiple sequence alignments. *Bioinformatics* **32**: 3501–3503.

Zuker M. 2003. Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res* **31**: 3406–3415.

Zuker M. 1989. On finding all suboptimal foldings of an RNA molecule. *Science* **244**: 48–52. http://www.ncbi.nlm.nih.gov/pubmed/2468181.

Tables

Table 1. Identification by RNAMotif and RNASurface of toxin-antitoxin mRNA-asRNA pairs corresponding to toxin ORF loci predicted by (Fozo et al. 2010).

TA family	# of ORFs predicted by (Fozo et al. 2010)	# of mRNA-asRNA pairs identified at the corresponding loci and with the corresponding family	# of mRNA-asRNA pairs identified at the corresponding loci but with an alternative family
Ldr/Rdl, Fst/RNAII[a]	299 (162, 136)	173 (136, 37)	16 (5, 11)
Hok/Sok	182	162	0
Ibs/Sib	210	182	0
ShoB/OhsC	31	26	3
TisB/IstR1	45	40	0
TxpA/RatA	122	88	12

[a]In (Fozo et al. 2010) members of the Ldr/Rdl and Fst/RNAII families were classified together due to remote similarity between toxin protein sequences of the two families (numbers in parentheses correspond to counts that would be obtained if systems from Gram-negative and Gram-positive organisms are assigned to the Ldr/Rdl and Fst/RNAII family, respectively).

Figure legends

Fig. 1. Automatic annotation of mRNAs and asRNAs of known type I TA systems.

Fig. 2. Screenshot of a "Locus Details" page in T1TAdb showing the various panels with sequence, structure, and genomic information.

Fig. 3. Structural alignments of two subgroups of Sok asRNAs from various Gram-negative bacteria. The structure of the subgroup shown in panel A matches that reported by (Franch et al. 1997) while the subgroup shown in panel B exhibits an additional stem-loop element at the 5' end. Alignments were computed using MAFFT (and slightly corrected manually) and annotated with covariation information using R-chie. A region of 30 nt upstream of the RNAs was added that includes matches to the -10 promoter box motif (UANNNU, where N means any nucleotide) highlighted in dark orange.

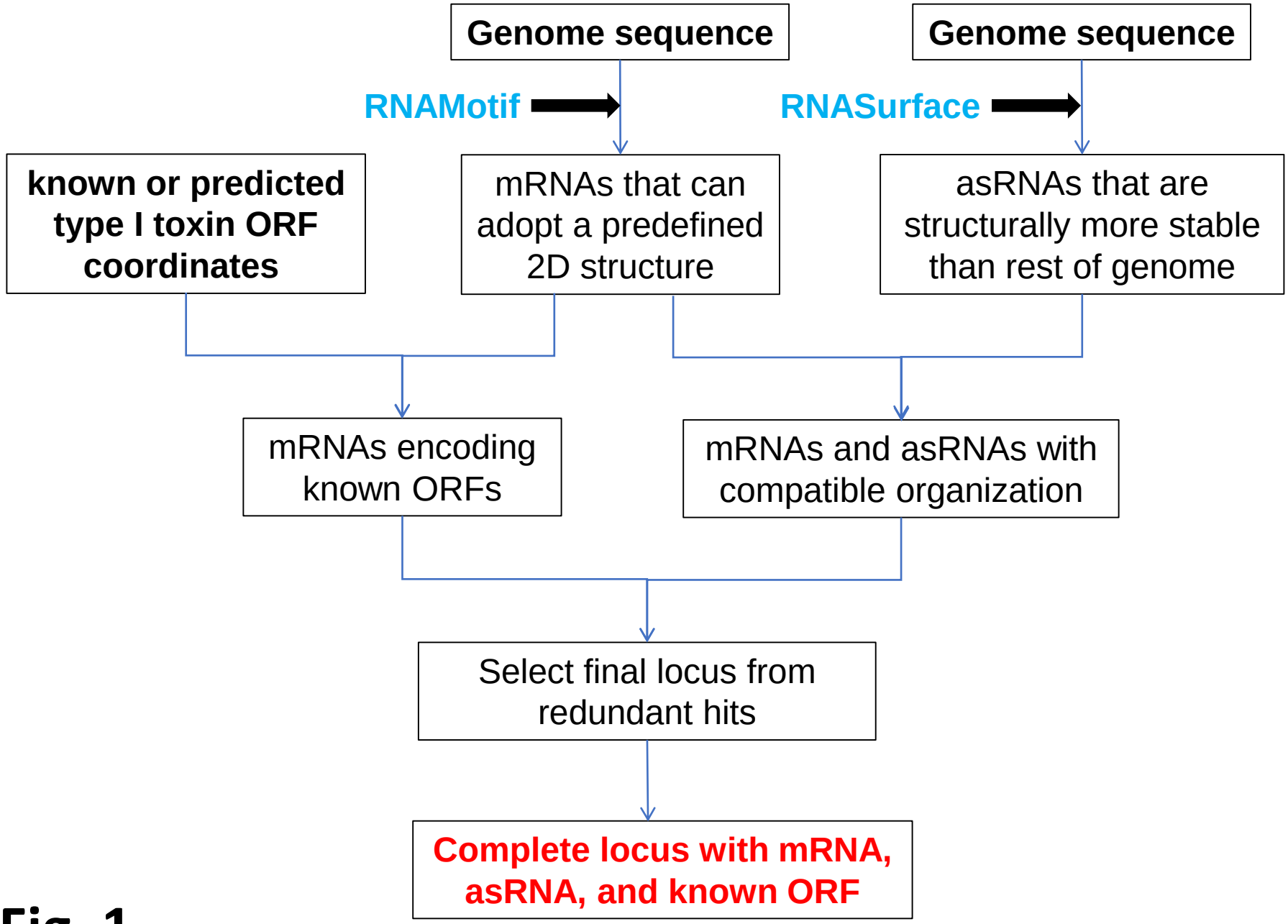

Genome sequence
Genome sequence
RNAMotif
RNASurface
known or predicted type I toxin ORF coordinates
mRNAs that can adopt a predefined 2D structure
asRNAs that are structurally more stable than rest of genome
mRNAs encoding known ORFs
mRNAs and asRNAs with compatible organization
Select final locus from redundant hits
Complete locus with mRNA, asRNA, and known ORF

Fig. 1

T1TAdb Locus Details

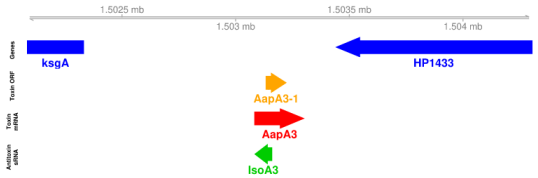
- Home
- Keyword Search
- Sequence Search
- Contact
- References
- Sponsors
- Links

Locus TA00126

Locus

ID	TA00126
Family	AapA/IsoA
Organism	Helicobacter pylori 26695 chromosome
Taxonomy	Bacteria; Proteobacteria; Epsilonproteobacteria; Campylobacterales; Helicobacteraceae; Helicobacter; Helicobacter pylori 26695
Genomic location	chromosome
Genomic coordinates	1503082-1503303
Genome accession (GenBank)	NC_000915.1
Genome map	CGView
Genomic context	IGV.js
Upstream gene	ksgA : rRNA small subunit methyltransferase A :
Downstream gene	HP1433 : hypothetical protein :

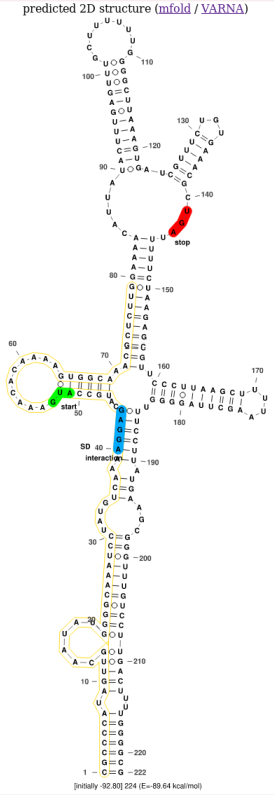
genomic context (Gviz / IGV.js)



Toxin mRNA

Name	AapA3
Locus_tag	
Genomic coordinates	1503082-1503303 (strand +)
Length	222 nt
Sequence (*with 50 nt upstream, and -10 box in red; mRNA/sRNA complementary region underlined; ORF in blue; transcript sequence in UPPERCASE*)	ttgatgcggtctaggggtgttttaagggggtttgttgtaggatttcatca CGCCCCATAGTTGCAATATGGGGCAAATCCTATGTCAAAAGGAGCATGCC ATGAAACACAAAAGTGGCAAACGCTCTTGGAAAACATTATACTTTGAGTT TGCTTTTTTGGGGCTTAAAGTGATCGTTTCTGTGAAACGCTGATTTTTCT AAGAGCGTTCCCTTAAGCTTTTAAGCTTAGGGGTTTCCTTATGAAGCGGG TTTGTCCTTGACTTTTGGGGCG → BLAST sequence
Predicted 2D structure	PDF, PNG

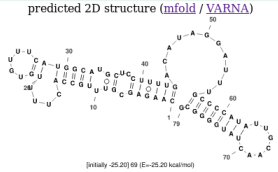
predicted 2D structure (mfold / VARNA)



Antitoxin asRNA

Name	IsoA3
Locus_tag	
Genomic coordinates	1503082-1503160 (strand -)
Length	79 nt
Sequence (*with 50 nt upstream, and -10 box in red; transcript sequence in UPPERCASE*)	aaacgatcactttaagccccaaaaaagcaaactcaaagtataatgttttc CAAGAGCGTTTGCCACTTTTGTGTTTCATGGCATGCTCCTTTTGACATAG GATTTGCCCCATATTGCAACTATGGGGCG → BLAST sequence
Predicted 2D structure	PDF, PNG

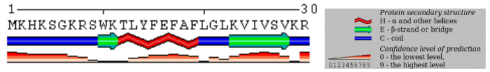
predicted 2D structure (mfold / VARNA)



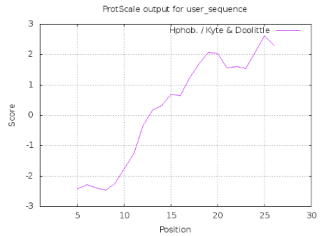
Toxin Peptide

Name	AapA3-1
Locus_tag	
ORF genomic coordinates	1503132-1503224 (strand +)
ORF length (peptide length)	93 nt (30 aa)
ORF sequence	ATGAAACACAAAAGTGGCAAACGCTCTTGGAAAACATTATACTTTGAGTT TGCTTTTTTGGGGCTTAAAGTGATCGTTTCTGTGAAACGCTGA → BLAST sequence
Peptide sequence	MKHKSGKRSWKTLYFEFAFLGLKVIVSVKR* → BLAST sequence
Predicted 2D structure	PDF, GIF
Hydrophobicity plot	PDF, GIF

predicted 2D structure (PSIPRED / POLYVIEW-2D)



hydrophobicity plot (ProtScale)



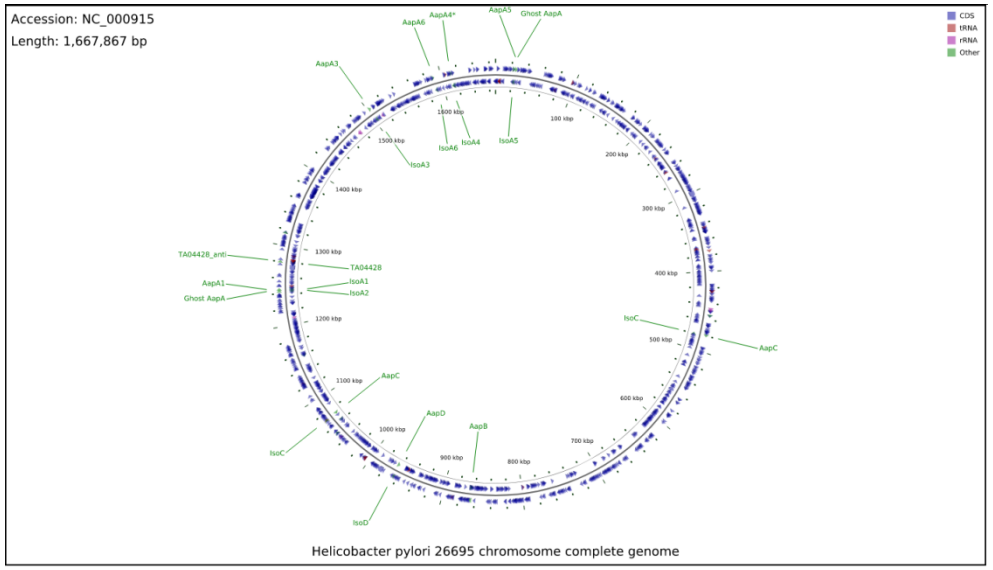


Fig. 2

Fig. 3A

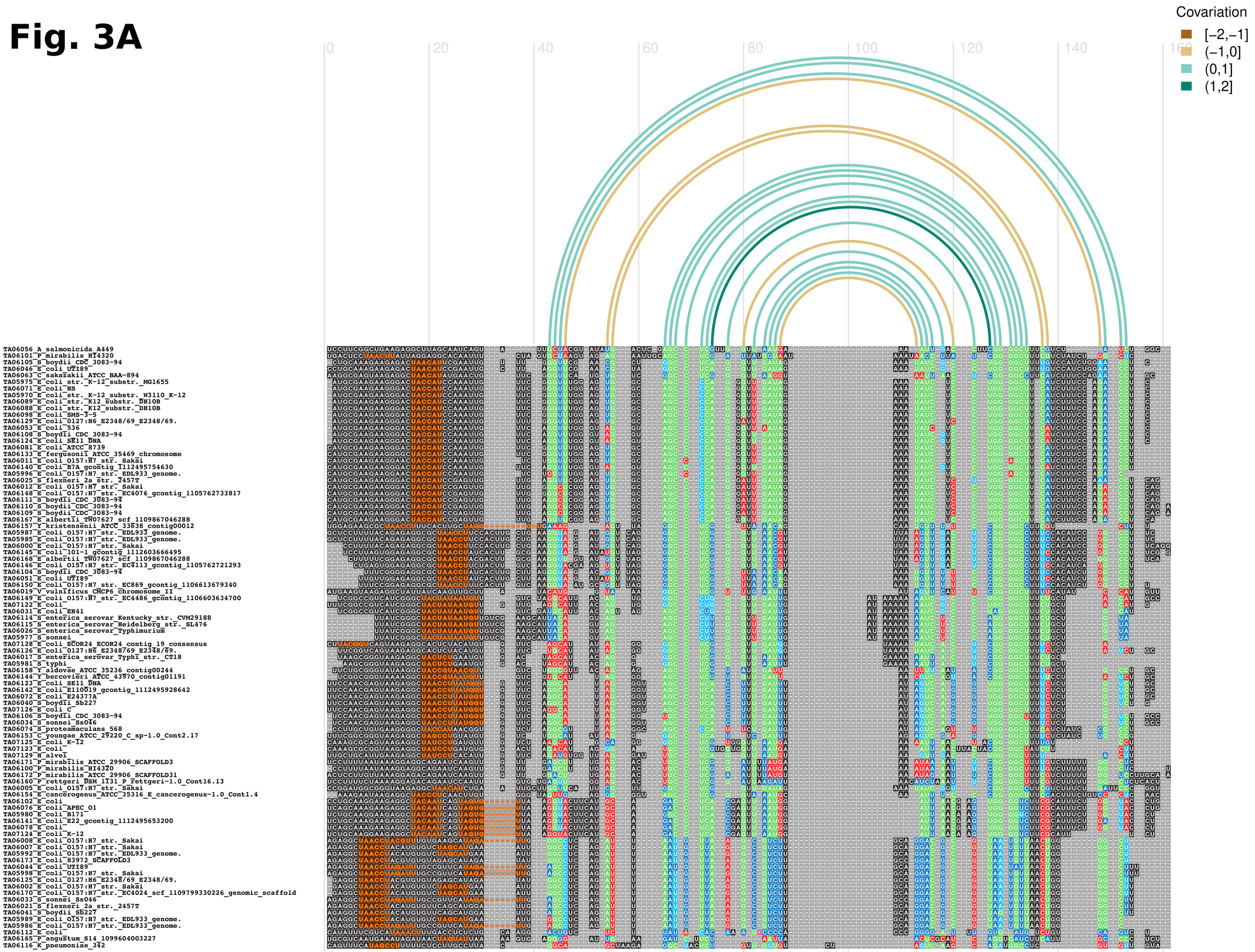
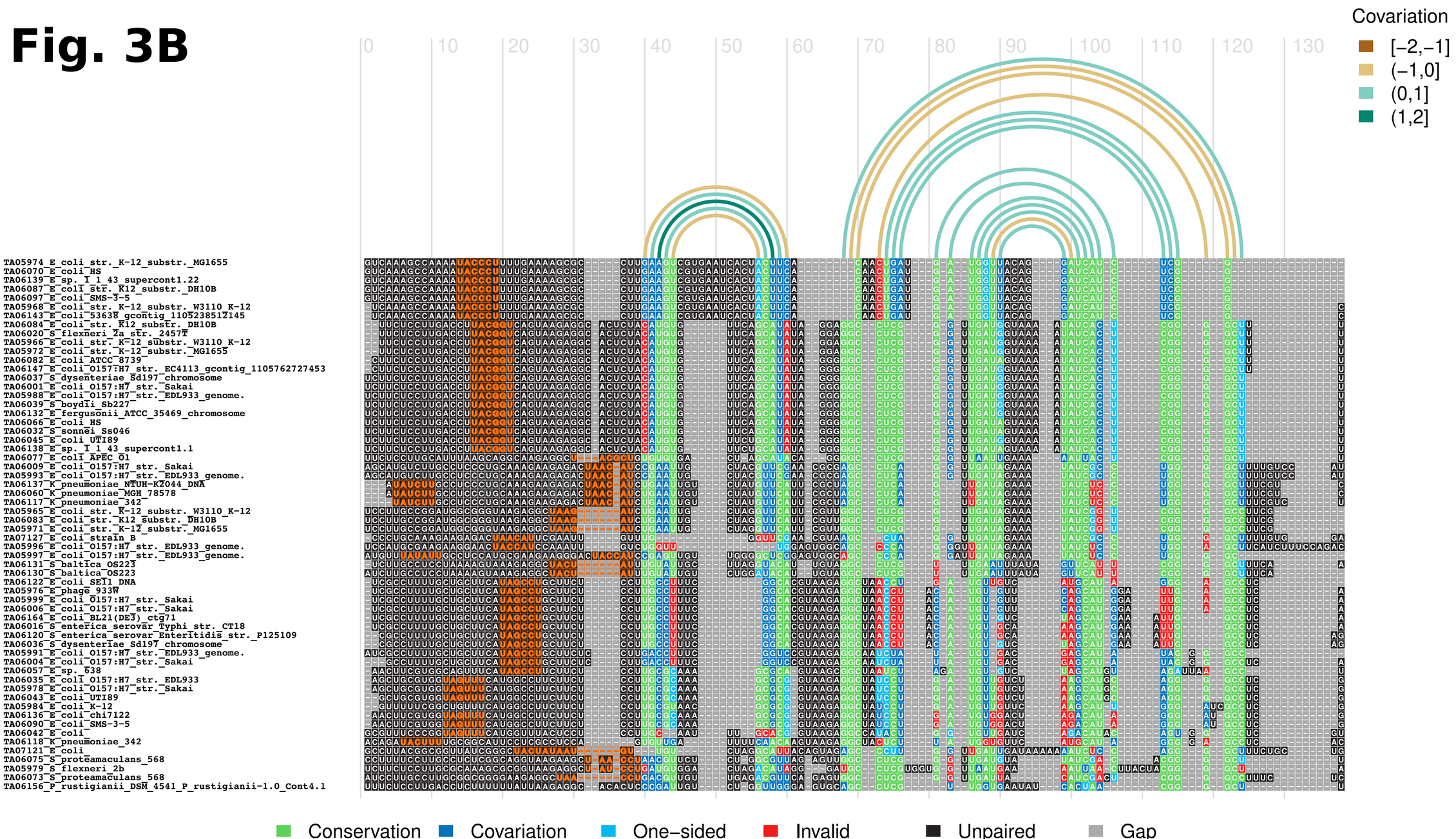

Covariation
[−2,−1]
(−1,0]
(0,1]
(1,2]
0
20
40
60
80
100
120
140
160
TA06056_A_salmonicida_A449
TA06101_P_mirabilis_HI4320
TA06105_S_boydii_CDC_3083-94
TA06046_E_coli_UMN189
TA06063_C_sakazakii_ATCC_BAA-894
TA05975_E_coli_str._K-12_substr._MG1655
TA06071_E_coli_HS
TA05970_E_coli_str._K-12_substr._W3110_K-12
TA06089_E_coli_str._K12_substr._DH10B
TA06088_E_coli_str._K12_substr._DH10B
TA06098_E_coli_SMS-3-5
TA06129_E_coli_O127:H6_E2348/69_E2348/69.
TA06053_E_coli_536
TA06108_S_boydii_CDC_3083-94
TA06124_E_coli_SE11_DNA
TA06081_E_coli_ATCC_8739
TA06133_E_fergusonii_ATCC_35469_chromosome
TA06011_E_coli_O157:H7_str._Sakai
TA06140_E_coli_B7A_gcontig_1112495754630
TA05996_E_coli_O157:H7_str._EDL933_genome.
TA06025_S_flexneri_2a_str._2457T
TA06012_E_coli_O157:H7_str._Sakai
TA06148_E_coli_O157:H7_str._EC4076_gcontig_1105762733817
TA06111_S_boydii_CDC_3083-94
TA06110_S_boydii_CDC_3083-94
TA06109_S_boydii_CDC_3083-94
TA06167_E_albertii_TW07627_scf_1109867046288
TA06157_Y_kristensenii_ATCC_33638_contig00012
TA05987_E_coli_O157:H7_str._EDL933_genome.
TA05985_E_coli_O157:H7_str._EDL933_genome.
TA06000_E_coli_O157:H7_str._Sakai
TA06145_E_coli_101-1_gcontig_1112603666495
TA06168_E_albertii_TW07627_scf_1109867046288
TA06146_E_coli_O157:H7_str._EC4113_gcontig_1105762721293
TA06104_S_boydii_CDC_3083-94
TA06051_E_coli_UTI89
TA06150_E_coli_O157:H7_str._EC869_gcontig_1106613679340
TA06019_V_vulnificus_CMCP6_Chromosome_II
TA06149_E_coli_O157:H7_str._EC4486_gcontig_1106603634700
TA07122_E_coli
TA06031_E_coli_EH41
TA06114_S_enterica_serovar_Kentucky_str._CVM29188
TA06115_S_enterica_serovar_Heidelberg_str._SL476
TA06026_S_enterica_serovar_Typhimurium
TA05977_S_sonnei
TA07128_E_coli_ECOR24_ECOR24_contig_19_consensus
TA06126_E_coli_O127:H6_E2348/69_E2348/69.
TA06017_S_enterica_serovar_Typhi_str._CT18
TA05981_S_typhi
TA06158_Y_aldovae_ATCC_35236_contig00244
TA06144_Y_bercovieri_ATCC_43970_contig01191
TA06123_E_coli_SE11_DNA
TA06142_E_coli_E110019_gcontig_1112495928642
TA06072_E_coli_E24377A
TA06040_S_boydii_Sb227
TA07126_E_coli_C
TA06106_S_boydii_CDC_3083-94
TA06034_S_sonnei_Ss046
TA06074_S_proteamaculans_568
TA06153_C_youngae_ATCC_29220_C_sp-1.0_Cont2.17
TA07125_E_coli_K-12
TA07123_E_coli
TA07129_H_alvei
TA06171_P_mirabilis_ATCC_29906_SCAFFOLD3
TA06100_P_mirabilis_HI4320
TA06172_P_mirabilis_ATCC_29906_SCAFFOLD31
TA06160_P_rettgeri_DSM_1131_P_rettgeri-1.0_Cont16.13
TA06005_E_coli_O157:H7_str._Sakai
TA06154_E_cancerogenus_ATCC_35316_E_cancerogenus-1.0_Cont1.4
TA06102_E_coli
TA06076_E_coli_APEC_O1
TA05980_E_coli_B171
TA06141_E_coli_E22_gcontig_1112495653200
TA06078_E_coli
TA07124_E_coli_K-12
TA06008_E_coli_O157:H7_str._Sakai
TA06007_E_coli_O157:H7_str._Sakai
TA05992_E_coli_O157:H7_str._EDL933_genome.
TA06173_E_coli_83972_SCAFFOLD3
TA06044_E_coli_UTI89
TA05998_E_coli_O157:H7_str._Sakai
TA06125_E_coli_O127:H6_E2348/69_E2348/69.
TA06002_E_coli_O157:H7_str._Sakai
TA06170_E_coli_O157:H7_str._EC4024_scf_1109799330226_genomic_scaffold
TA06033_S_sonnei_Ss046
TA06021_S_flexneri_2a_str._2457T
TA06041_S_boydii_Sb227
TA05989_E_coli_O157:H7_str._EDL933_genome.
TA05986_E_coli_O157:H7_str._EDL933_genome.
TA06112_E_coli
TA06165_P_angustum_S14_1099604003227
TA06116_K_pneumoniae_342
Conservation
Covariation
One-sided
Invalid
Unpaired
Gap
Ambiguous

Fig. 3B



www.e−rna.org