A vast collection of microbial genes that are toxic to bacteria

Aya Kimelman[1†], Asaf Levy[1†], Hila Sberro[1†], Shahar Kidron[1], Azita Leavitt[1], Gil Amitai[1], Deborah R. Yoder-Himes[2], Omri Wurtzel[1], Yiwen Zhu[3,4], Edward M Rubin[3,4], Rotem Sorek[1,*]

Supplementary Material

Supplementary tables

Supplementary Figures

Supplementary data files

Supplementary Methods

Identification of toxic genes using clone coverage analysis

The raw paired-end sequencing reads (mate-pairs) for each processed genome were downloaded from NCBI Trace Archive (http://www.ncbi.nlm.nih.gov/Traces/) in multi-fasta format, along with ancillary information describing trace information for each read. In genome project for which traces were available, the corresponding finished genomes were downloaded from the NCBI RefSeq database (http://www.ncbi.nlm.nih.gov/RefSeq/), or NCBI GenBank database (http://www.ncbi.nlm.nih.gov/genbank/). Genomes were further analyzed only if their sequencing status was determined as "finished" or "completed" in the Genomes Online Database (GOLD; http://www.genomesonline.org/). Supplementary gene annotations, such as COG and TIGRFAM IDs, were downloaded from the IMG v3.2 database (http://img.jgi.doe.gov/cgi-bin/w/main.cgi), if existed.

Each batch of sequencing traces was initially processed to discard traces that applied to non clone-based sequencing methods (such as 454 and Illumina), traces that were considered obsolete (i.e have been replaced by an updated trace or have been withdrawn), and traces that lacked information regarding direction of read (forward or reverse corresponding to the 5' or 3' end of the clone, respectively). The fasta sequences of the remaining clone based reads were trimmed according to clip information available in the Trace Archive, to remove vector and/or poor quality sequences.

The trimmed read sequences were mapped to their corresponding reference genome as previously described[1]. Briefly, the NUCmer application from the MUMmer3.21 software package (http://mummer.sourceforge.net/) was used with the <-maxmatch> parameter, thus computing all maximal matches regardless of their uniqueness. The delta-filter application from the MUMmer3.21 software package was then used with the <–i> and <–l> parameters, to discard reads with less than 95% identity or less than 300 aligned bases to the reference genome, considering that reads produced by clone based sequencing methods, such as Sanger sequencing, vary between 600-1000 bp in length. Based on read alignments, clone positions on the reference genome were inferred from the mapping of both mate reads (representing the 5` and 3` sequenced ends pertaining to the same clone). In ambiguous cases, where a read mapped to several positions on the reference genome, thus producing several possible clones with its mate read, the correct position was inferred by the NCBI ancillary trace information indicating the sequenced clone expected length.

Clones were considered invalid and discarded from further consideration if both mate reads mapped to the same strand, to different replicons (chromosomes/plasmids), or produced a clone size that did not coincide with the trace information indicating mean library insert size and standard deviation (clones were considered valid if they did not exceed a distance of 2.5 standard deviations from the defined library mean insert size). In the absence of insert size information, an estimation of the mean insert size and standard deviation was computed based on actual clone lengths produced by paired reads of the same annotated library or sequencing plate, which mapped uniquely to the reference genome to produce a valid clone. If no such library or plate information existed, or if there were not enough uniquely mapped valid clones to assume normal distribution of clone lengths (N<50), all clones shorter than 50kb were considered as valid. After discarding all invalid clones, there were cases in which some pairs of mate reads produced more than one valid mapping to the genome. In such cases, one of the mappings was selected randomly per each mate pair.

Per-base coverage was determined for each nucleotide position in the reference genome, by counting the number of valid mapped clones that spanned that position. Replicons for which the average clone coverage depth was less than 10x were considered as insufficiently covered, and did not participate in any further coverage analysis. Furthermore, replicons in which more than 3% of the genomic sequence was not covered by a single clone were considered to be outliers and were ignored as well.

A flowchart of the data analysis pipeline is visually presented in Fig S2.

A statistical framework to assess gene clonability

Gene positions were defined according to the "gene" feature annotation in the reference genomes' GenBank files. For each gene, the number of clones fully spanning it was recorded. In order to define if a gene was covered by fewer clones than expected by chance, we developed a statistical framework to quantify "unclonability" of a gene and assign it with a p-value. Such quantification allowed comparisons between genes across replicons and genomes that vary in coverage depths. Furthermore, such statistical quantification is important as long genes are less plausible to be fully covered by a single clone, thus their low clone coverage does not necessarily imply any biological effect.

To assess the significance of low or zero-coverage of a gene, 100 coverage simulations were performed, by randomly shuffling the positions of all valid sequencing clones on the reference genome. The number of clones covering each gene was obtained per simulation, and the mean of random clone coverage was calculated per gene (N=100). A p-value for the actual gene clone coverage was calculated, relative to a cumulative Poisson distribution of the random coverage values (λ=mean), based on an observed fit between the results of the simulations and the Poisson distribution. P-values for each gene were then corrected for multiple hypothesis testing using FDR correction (N=number of genes per genome).

Additional processing was then performed to identify "hitchhiker" genes. Hitchhiker genes are thought to be genes whose unclonability is likely to be due to a near-by genomic element, such as a neighbor gene, which is the core reason for the unclonability of sequencing clones spanning that area. Such hitchhiker genes were detected by searching for the local minimum of nucleotide clone-coverage in a window surrounding the gene in question. The window size used was the median of all clones mapped to the genome. If the lowest nucleotide clone-coverage value was not within the gene coordinates, the gene was marked as a "hitchhiker".

Eventually, each gene was assigned with an unclonability value ("unclonable", "decreased coverage", "hitchhiker" , "normal" , “n/a”) according to the number of clones fully containing it, the assigned p-value indicating the probability of obtaining such coverage by chance, and the hitchhiker analysis. A gene was considered to have significantly low clone coverage if $p < 0.01$ after correction for multiple testing.

Initial experimental evaluation of gene toxicity

Genes were selected for toxicity experiments based on the following criteria: 1. the genomic DNA of the relevant bacteria was available at the lab in quantities enabling PCR amplification of DNA template; 2. the function of the selected gene was unknown or only generally annotated; 3. preference was given for short genes to increase chances of PCR amplification. Cloning of genes under the control of an inducible promoter was performed as previously described [1]. Briefly, genes were amplified from their genome of origin and directionally cloned into the pET-11a vector (Stratagene Santa Clara, CA, USA) within *E. coli* BL21- Gold(DE3)pLysS cells (Stratagene). For the expression activation experiment, clones were cultured in LB medium with 100 μg/ml ampicillin and 34 μg/ml chloramphenicol overnight. The next day, a portion of each overnight culture was inoculated into fresh medium (20-fold dilution) and cultured at 37C for 2 to 3 hours with 250rpm shaking. Cells were then diluted 2500-fold, and 10 microliters of diluted cells of each culture were spotted into a well of 48-well plates containing LB agar with or without 100, 250, 400, 600 and 800 μM of IPTG. The plates were incubated overnight at 37 °C, and growth of each clone in the different IPTG conditions was recorded. Genes were considered "very toxic" if IPTG concentrations of 100 μM were sufficient to eliminate growth; "toxic" if IPTG concentrations of 250-400 μM were sufficient to eliminate growth; "mildly toxic" for IPTG conc. of 600 μM; and "weakly toxic" for IPTG conc. of 800 μM. As negative control, 15 clonable genes were used[1].

Prediction of toxic small RNAs within uncloned intergenic regions

Refseq/Genbank replicons of sufficient clone coverage were scanned for intergenic sequences of zero plasmid coverage that are flanked by genes covered by at least one clone. Intergenic sequences that were longer than 1500 bp were filtered out as well as DNA sequences containing stretches of Ns (unknown nucleotide). This screen resulted in a list of 873 intergenic sequences of average size 430 bp found within 274 distinct replicons (Table S4).

The next step was to try to assign functionality to these unclonable intergenic sequences. Since the 16S ribosomal RNA promoter is known to be toxic when cloned into plasmids[2], the 114 intergenic regions located next to the 16S gene were flagged and discarded (most of these sequences (92%, 105/114) were promoters of rRNAs from *Beta-* or *Gammaproteobacteria*). ORFs were then predicted within the remaining 759 intergenic sequences using the getorf script from the EMBOSS package[3], allowing only ORFs larger than 50 amino acids, bearing an ATG start codon, and terminated by a valid stop codon. Conserved ORFs were searched by blastp against the NCBI nr database with an e-value threshold of 0.05. A conserved ORF was defined as an ORF that has at least 2 hits against nr proteins. Eventually 37 intergenic sequences were annotated as containing at least one conserved ORF (41 conserved ORFs in total).

Next, we searched for small noncoding RNAs (sRNAs/ncRNAs). For this purpose we scanned the remaining 722 unannotated intergenic sequences for a conservative combination of a promoter, Rho-independent transcription terminator, a lack of an internal ORF, size of at least 20 bases, and sequence conservation. Promoters were identified based on RpoD and RpoS sigma factors binding sites of different spacer sizes between -35 and -10 boxes (taken from http://arep.med.harvard.edu/ecoli_matrices/). Position specific scoring matrices for the sigma factor binding sites were constructed and were used to scan the intergenic sequences using EMBOSS prophecy and profit scripts, respectively[3]. Promoters were assigned to 696 of the intergenic sequences. Terminators were identified using TranstermHP program[4]. Intergenic terminators that were located immediately downstream (up to 50 bp) to the flanking genes were discarded. Terminators were assigned to 301 of the intergenic sequences. Predicted transcripts were defined as a combination of promoter and terminator within the same orientation, with the promoter being upstream to the terminator, and the transcript being longer than 20 bases. Transcription start site was defined as seven bp downstream to the promoter position and transcription end was defined as the end of the consecutive terminal poly-U stretch that follows the terminator hairpin. Transcripts were predicted for 244 intergenic sequences with several cases of multiple non-overlapping transcripts derived from the same intergenic region (in total, 315 transcripts were predicted). In the next step, we further filtered out previously unidentified ORFs within the predicted transcripts. This ORF group included ORFs characterized by length longer than 30 amino acids, and a start codon that corresponds to ATG, GTG or TTG, without requiring similarity to known proteins. However, in order to compensate for these relaxed criteria, we required a putative ribosome binding site (RBS) upstream to the start codon (defined as at least four consecutive purines within positions -13 to -7 to the start codon). The second group of ORFs contained nine more ORFs that sum up to 50 ORFs in total from 46 intergenic sequences, together with the conserved ORFs (File S1). From the remaining putative noncoding transcripts we further filtered four sRNA candidates that were adjacent to the previously predicted ORFs. The resulting list contains 302 sRNA candidates (File S2; Table S4).

To identify the sRNAs with reproducible unclonability all candidate sequences were clustered together based on homology (at least 80% identity along 70% of the predicted transcript length) using blastclust program[5]. Clusters that were composed of only sRNA candidates from the same intergenic region were filtered out. Clustering was further refined based on homology between the genes flanking the intergenic sequences (synteny). The clustering step yielded a list of 17 sRNA candidate clusters (Table S5) containing 69 sRNA candidates in total.

Motif enrichment analysis and dnaA boxes

Intergenic regions in which no ncRNAs, ORFs, or rRNA operon promoters were detected, were searched for overrepresented motifs by scanning the sequence using the wordcount program from the emboss package[3] using different wordsizes. For a wordsize of 9 bp we identified the sequence TTATCCACA as the most abundant sequence (65 occurences) and its antisense sequence as the 2nd most abundant sequence (61 occurences). In order to check whether this sequence is indeed over-represented within unclonable sequences we compared the expected number of occurrences of this sequence to the observed number of occurrences using Fisher exact test. The expected number is the multiplication of the occurrence of the 9-mer within all replicons in our dataset (n=12866) by the fraction of the intergenic unclonable sequences out of all replicon sequences (0.00014). The resulting expected number of this 9-mer within the unclonable sequences is 1.85. Fisher exact test results were $p = 3.6 \times 10^{-35}$ and odds ratio = 64.

The DnaA box consensus sequence (TTATCCACA) with maximum two mismatches was searched within all 768 Refseq replicons of sufficient plasmid coverage. The search was committed using a Perl script and the dreg program from the EMBOSS package[3]. DnaA boxes were clustered together if they were up to 20 bp apart from each other. Namely, every adjacent DnaA boxes within a DnaA box cluster are located within 20 bp at most from each other. Number of plasmid clones covering each DnaA box/cluster was calculated based on genomic positions of both the DnaA box/cluster loci and clones as calculated above. Unclonability of a DnaA box/cluster was defined as being covered by a zero plasmid clones. As a control, 60 random sequence shufflings of DnaA box consensus sequence (after checking that shuffled sequences have more than two mismatches) and 60 random 9-mer bp sequences were created. Unclonability was checked for the control sequences using the same criteria as above. The result was a set of 57 shuffled DnaA box clusters and 58 random 9-mer bp clusters. Visualization of the results was done using the Artemis genome browser[6] and R programming language. The datA locus sequence within *E. coli* HS (NC_009800) was identified based on the literature[7]. The orthologous datA loci were identified by blastn and unclonability was determined as above.

Supplementary figures

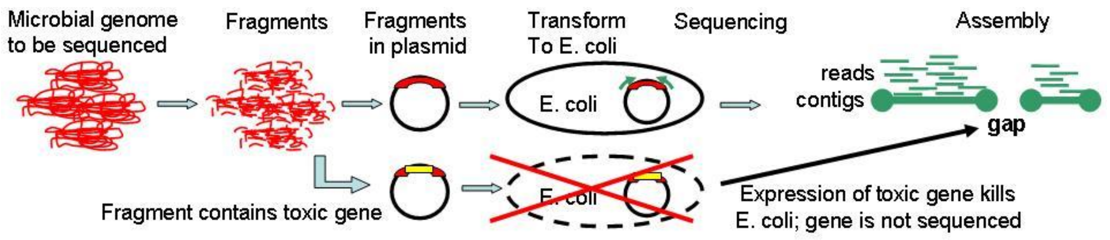
Microbial genome to be sequenced
Fragments
Fragments in plasmid
Transform To E. coli
Sequencing
Assembly
E. coli
reads
contigs
gap
Fragment contains toxic gene
E. coli
Expression of toxic gene kills E. coli; gene is not sequenced

Fig S1: The microbial genome sequencing process identifies genes that inhibit *E. coli* growth. Multiple copies of the genome to be sequenced are physically sheared into overlapping fragments of DNA (typically sized 3kb, 8kb or 35kb). Cloned fragments are transformed into *E. coli* cells on plasmids. Resulting sequence reads are assembled into larger contigs. In some genomes a small fraction of the organism's genome fails to clone in *E. coli*, resulting in sequence gaps. The sequence for these uncloned gaps is acquired via a "finishing" stage that eventually produces an unbroken, continuous sequence of the genome[8]. Such gaps can occur because of genes whose products are toxic to the *E. coli* host. When these genes are cloned into *E. coli*, their expression product inhibits bacterial growth, and hence the inability to clone and sequence them.

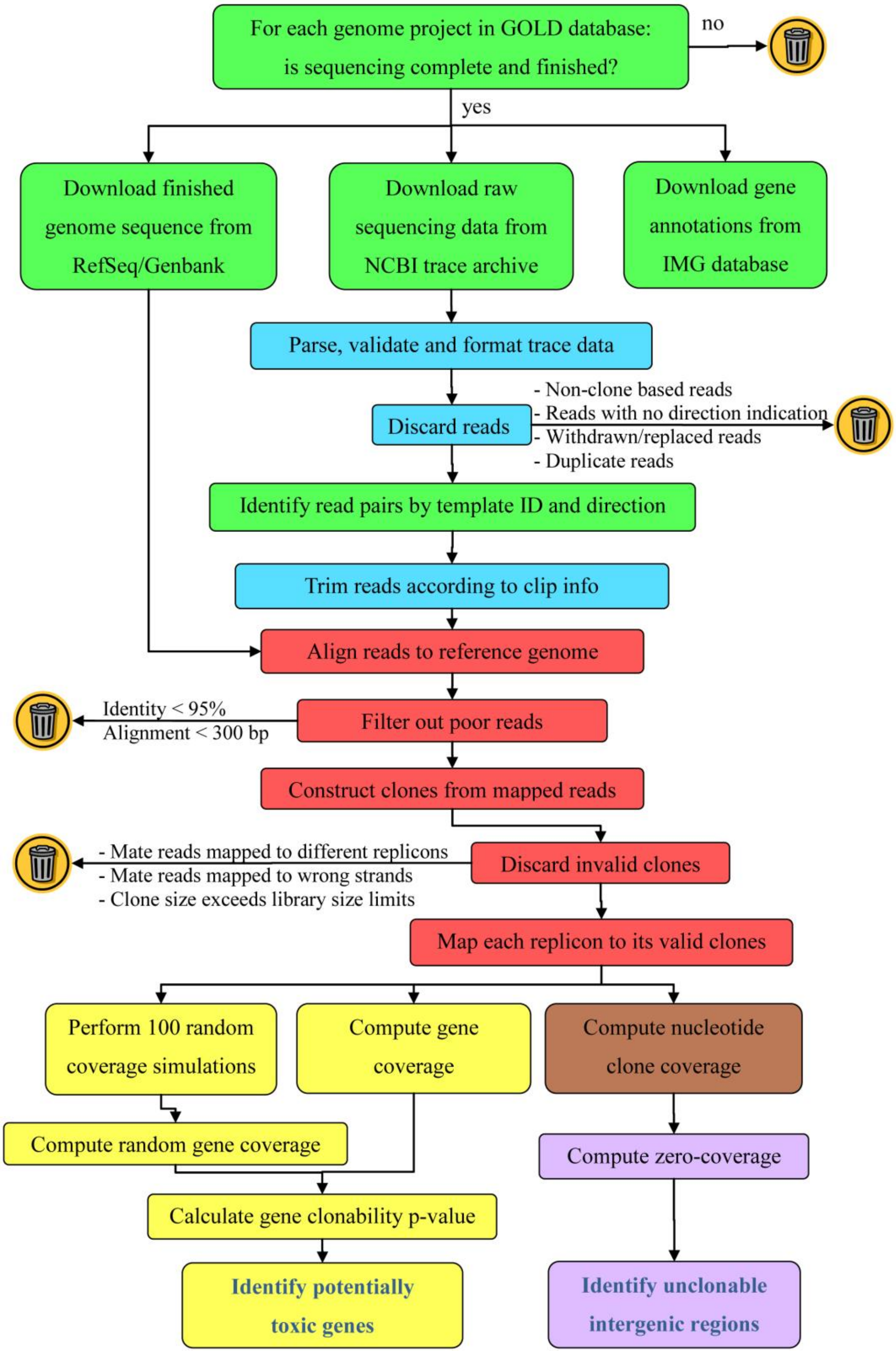
For each genome project in GOLD database: is sequencing complete and finished?
no
yes
Download finished genome sequence from RefSeq/Genbank
Download raw sequencing data from NCBI trace archive
Download gene annotations from IMG database
Parse, validate and format trace data
Discard reads
- Non-clone based reads
- Reads with no direction indication
- Withdrawn/replaced reads
- Duplicate reads
Identify read pairs by template ID and direction
Trim reads according to clip info
Align reads to reference genome
Identity < 95%
Alignment < 300 bp
Filter out poor reads
Construct clones from mapped reads
- Mate reads mapped to different replicons
- Mate reads mapped to wrong strands
- Clone size exceeds library size limits
Discard invalid clones
Map each replicon to its valid clones
Perform 100 random coverage simulations
Compute gene coverage
Compute nucleotide clone coverage
Compute random gene coverage
Compute zero-coverage
Calculate gene clonability p-value
Identify potentially toxic genes
Identify unclonable intergenic regions

Fig S2: Data analysis flowchart for the identification of toxic genes.

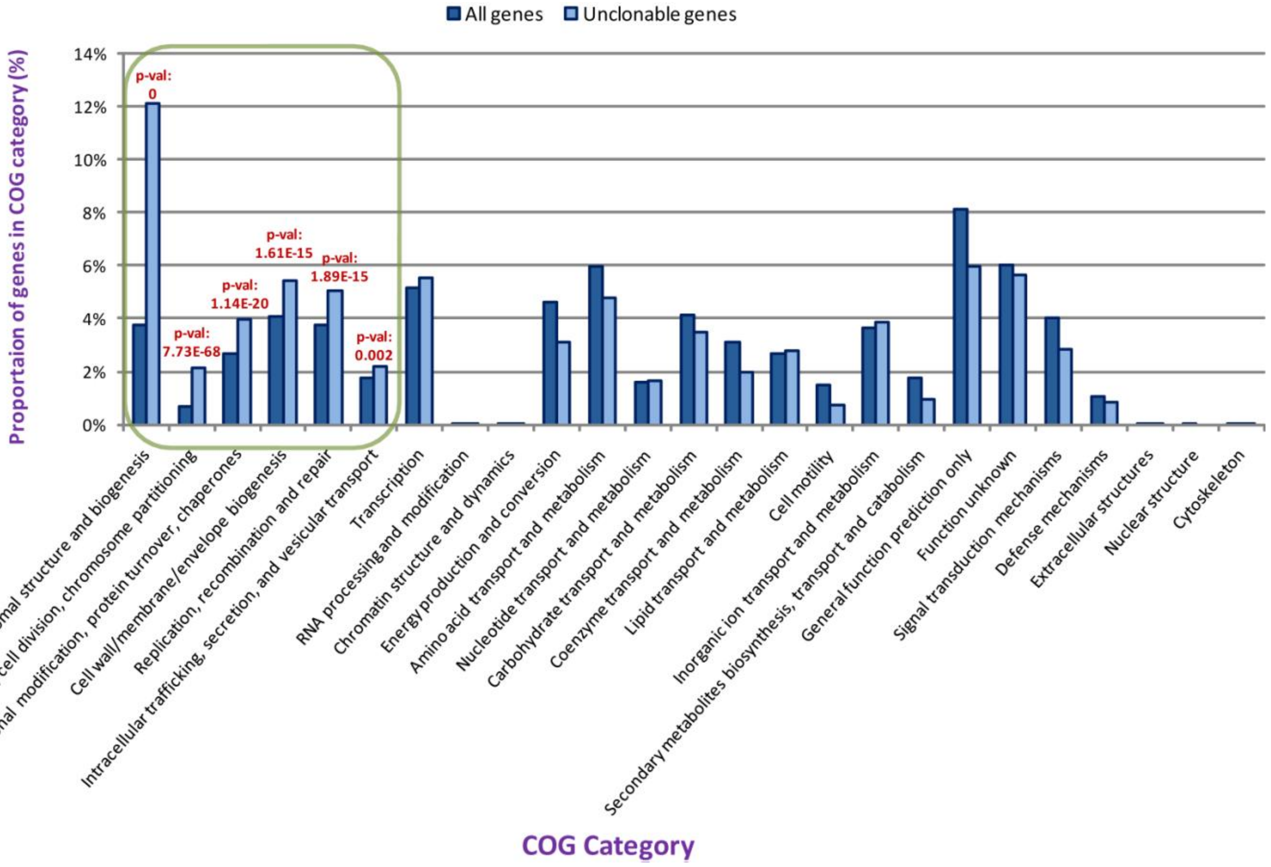

Fig S3: COG category enrichment within all annotated unclonable genes as compared to clonable genes.

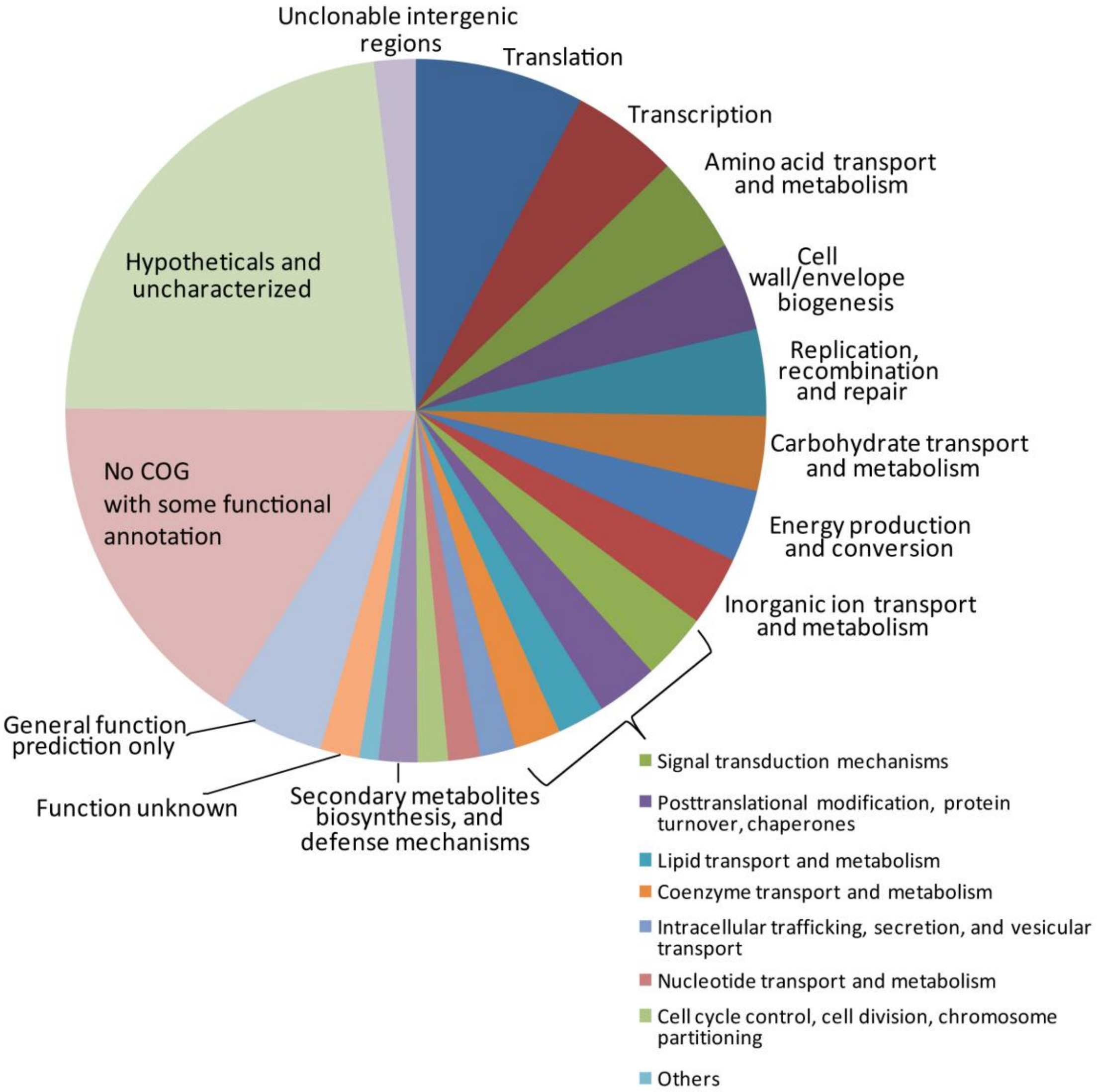

Fig S4: Functional distribution of all toxic genes in the PanDaTox database

Fig S5 -A

```
NC_007434_37401 cattggttgtcgcgttg------------------cggcaacctccgca--------tgtctcctccac-
NC_009076_33308 cattggttgtcgcgttg------------------cggcaacctccgca--------tgtctcctccac-
NC_008390_64585 cattggttgtcgcggtg------------------cgacaatctccgca--------tgtctcctccac-
NC_010551_67662 cattggttgtcgcggtg------------------cgacaatctccgca--------tgtctcctccac-
NC_010622_28852 -attggttgtcgaact-------------------cgacaatctccgca--------tgtctcctccac-
NC_010681_41308 catgttgcgccgcacca------------------agccgtttgctatagtttgtcttgtctcctccatg
NC_007951_45704 ---------------------------------------------------------tgtctcctccatg
NC_008542_35198 ---------------------------------------------------------tgtctcctccatg
NC_010508_31993 ---------------------------------------------------------tgtctcctccatg
NC_010551_27317 ---------------------------------------------------------tgtctcctccatg
NC_008060_30746 ---------------------------------------------------------tgtctcctccatg
NC_007951_48755 ------gcgtcgcacca------------------accccatgtgtatattagtacttgtctcctccatg
NC_010622_34684 ---------------------------------------cgcgtgtatagtggttcctgtctcctccatg
NC_008060_27916 ------------cacca------------------tcgccgtgtgtatagttggaactgtctcctccatg
NC_010508_35225 ------------cacca------------------tcgccgtgtgtatagttggaactgtctcctccatg
NC_008390_35460 ---------------------------------------------------------tgtctcctccatg
NC_010551_34333 ---------------------------------------------------------tgtctcctccatg
NC_008785_28178 ------------cacca------------------tcccgatgtgtatagtgggaactgtctcctccatg
NC_008836_22771 ------------cacca------------------tcccgatgtgtatagtgggaactgtctcctccatg
NC_009080_22668 ------------cacca------------------tcccgatgtgtatagtgggaactgtctcctccatg
NC_009080_22853 ------------cacca------------------tcccgatgtgtatagtgggaactgtctcctccatg
NC_007651_89266 ------------------------------------------gtgtatagttcgaactgtctcctccatg
NC_003295_52707 cattggatcggctgacacggtgcaacccactgggtcaacgatcagcagg--------tgtctcctccac-
NC_010682_39135 cattggatcggctgatgcggtgcaacccaccaagtcaacgatcagcagg--------tgtctcctccac-
                                                                         ***********

NC_007434_37401 cctcctcctttggtggattaag--cccgaacc-----agcggttcgggcttttttt-
NC_009076_33308 cctcctcctttggtggattaag--cccgaacc-----agcggttcgggcttttttt-
NC_008390_64585 cctcctcctgaggtggattaag--cccgaacc-----agccgttcgggcttttttt-
NC_010551_67662 cctcctcctgaggtggattaag--cccgaacc-----agccgttcgggcttttttt-
NC_010622_28852 cctcctcctaaggtggattaag--cccgaacc-----gctagttcgggcttttttt-
NC_010681_41308 tctcct-ctgatatggattcag--cccgcca------aattaggcgggctttttttt
NC_007951_45704 tctcct-ctgatatggattcag--cccgccc------aaataggcgggctttttttt
NC_008542_35198 tctcct-ctgatatggattcag--cccgcc-------acttaggcgggctttttttt
NC_010508_31993 tctcct-ctgatatggattcag--cccgcc-------acttaggcgggctttttttt
NC_010551_27317 tctcct-ctgatatggattcag--cccgcc-------tcttaggcgggctttttttt
NC_008060_30746 tctcct-ctgatatggattcag--cccgcc-------acttaggcgggcttttttt-
NC_007951_48755 tctcctcctgatatggattcag--cccgctcca----catagagcgggcttttttt-
NC_010622_34684 tctcctcctgatatggattcag--cccgctcca----cgtcgagcgggctttttttt
NC_008060_27916 tctcctcctgatatggattaag--cctgttccgctctgcgtgaacgggcttttttt-
NC_010508_35225 tctcctcctgatatggattaag--cccgttccgctctgcgtgaacgggcttttttt-
NC_008390_35460 tctcctcctgatatggattaag--cccgttccgctctgcgtgaacgggcttttttt-
NC_010551_34333 tctcctcctgatatggattaag--cccgttccgctctgcgtgaacgggcttttttt-
NC_008785_28178 tctcctcctgatatggattaag--cccgttcga----acatgaacgggcttttttt-
NC_008836_22771 tctcctcctgatatggattaag--cccgttcga----acatgaacgggcttttttt-
NC_009080_22668 tctcctcctgatatggattaag--cccgttcga----acatgaacgggcttttttt-
NC_009080_22853 tctcctcctgatatggattaag--cccgttcga----acatgaacgggcttttttt-
NC_007651_89266 tctcctcctgatatggattaag--cccgttcga----atgtgaacgggcttttttt-
NC_003295_52707 cctcctcctttggtggattcaaaccccaagcca----aaccgcttggggtttttttt
NC_010682_39135 cctcctcctttggtggattcaaa-cccaagctt----gatggcttgggtttttttt-
                 ***** ** : .******.*.    ** .           .   *** *******
```

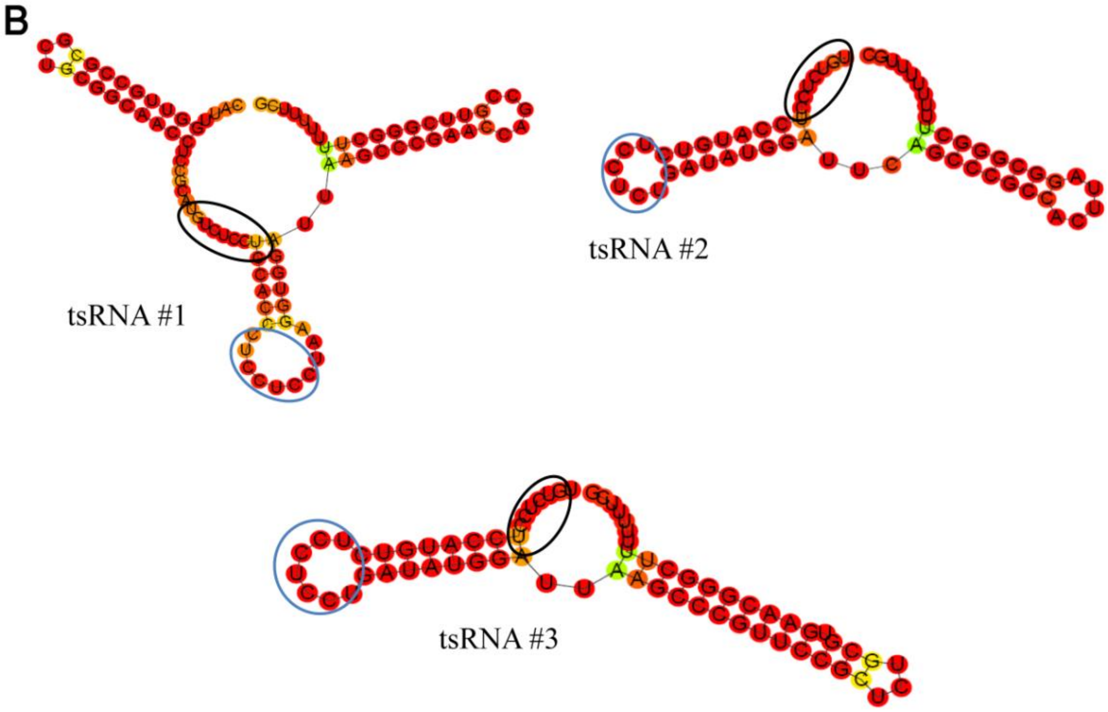

Figure S5. Conserved motif in tsRNAs may target bacterial ribosomal binding sites (RBSs) (A) A multiple sequence alignment of 31 sequences of tsRNAs 1-3. A dashed black rectangle denotes a highly conserved 11mer motif, which harbors a sequence complementary to the consensus AGGAGA RBS. A blue solid rectangle denotes a partially conserved pyrimidine-rich motif complementary to RBS sequences. Red highlighted text denoted structurally-open motifs that are expected to target RBS sequences (AGGAGG, AGGAGA, or AGAGGA) of target genes. Alignment was performed using MAFFT[9]. (B) Predicted RNA secondary structure of the three tsRNAs from Burkholderia *cenocepacia* HI2424. Black and blue ellipses denote the two putative anti-RBS motifs presented in panel A. These motifs are shown to reside in a predicted open (unfolded) region in all tsRNAs. Folding was predicted using RNAfold[10].

Supplementary References

[1] Sorek, R. et al., Genome-wide experimental determination of barriers to horizontal gene transfer. *Science* 318 (5855), 1449 (2007).

[2] Boros, I. et al., Cloning of the promoters of an Escherichia coli rRNA gene. New experimental system to study the regulation of rRNA transcription. *Gene* 22 (2-3), 191 (1983).

[3] Rice, P., Longden, I., and Bleasby, A., EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet* 16 (6), 276 (2000).

[4] Kingsford, C. L., Ayanbule, K., and Salzberg, S. L., Rapid, accurate, computational discovery of Rho-independent transcription terminators illuminates their relationship to DNA uptake. *Genome Biol* 8 (2), R22 (2007).

[5] Altschul, S. F. et al., Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 25 (17), 3389 (1997).

[6] Rutherford, K. et al., Artemis: sequence visualization and annotation. *Bioinformatics* 16 (10), 944 (2000).

[7] Kitagawa, R., Mitsuki, H., Okazaki, T., and Ogawa, T., A novel DnaA protein-binding site at 94.7 min on the Escherichia coli chromosome. *Mol Microbiol* 19 (5), 1137 (1996).

[8] Gordon, D., Desmarais, C., and Green, P. Automated finishing with autofinish. *Genome Res* **11:** 61 (2001).

[9] Katoh, K., Misawa, K., Kuma, K., and Miyata, T., MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res* 30 (14), 3059 (2002).

[10] Hofacker, I. L. et al., Fast Folding and Comparison of Rna Secondary Structures. *Monatshefte Fur Chemie* 125 (2), 167 (1994).