1. Supplementary Text and Figure Legends.

S1 PDF file (tracRNA Characteristics.pdf) shows the statistics of known tracrRNAs: (A) The distribution of Cas9-tracrRNA intergenic distances. (B) Conserved nucleotides at the start of CRISPR repeats. (C) Conserved nucleotides at the start of the tracrRNA tails. (D) For known tracrRNAs the distribution of repeat-anti-repeat free-energy calculated by RNAhybrid.

S2 PDF file (SVM testing.pdf) shows the SVM testing statistics: (A) The distributions of ‘length of the longest helix’ plotted against the proportions of non-pairing nucleotides in repeat-anti-repeat hybrids of known tracrRNAs (triangles) and randomly-formed hybrids (circles). Both the length of the longest helix and the proportion of non-pairing in repeat-anti-repeat hybrids are good discriminators. (B) Leave-one-out cross-validation of the SVM used. (C) the distribution of the SVM true-positive probabilities for known tracrRNAs, only one falls below the cutoff used (0.970).

# S3 CSV-formatted table (tracrRNA all prediction.csv) shows all tracrRNA predictions within 3500 bases of a *cas9* gene and with a SVM true-positive probability above 0.98. The explanations of the columns (column name (MS Excel column index)) are:

# ORGANISM (A) Name of the organism.

# CONTIG\_ID (B) ID of the contig with tracrRNA.

# RR\_CONTIG (C) ID of the contig with repeat array (tracrRNA and repeat array may be on different contigs).

# RR\_ID (D) ID of the repeat array (appended with \_RC if it is the reverse complement of the initial prediction).

# CD\_SCORE (E) Score (CRISPRDetect) of the repeat array predicted.

# REPEAT\_SEQ (F) The repeat sequence from CRISPRDetect.

# RR\_START (G) The start-coordinate of the repeat array calculated by CRISPRDetect.

# RR\_END (H) The end-coordinate of the repeat array calculated by CRISPRDetect.

# RR\_ORI (I) Orientation of the repeat array calculated by CRISPRDetect (1 if it is on the forward strand , 0 otherwise).

# CAS9\_INFO (J) A list of annotated Cas-family proteins (taken from CRISPRDetect TXT output).

# BLAST\_HSP\_START (K) The start-coordinate of the BLASTN alignment between the contig and the repeat sequence (i.e. the hot spot, HSP).

# BLAST\_HSP\_END (L) The end-coordinate of the BLASTN alignment between the contig and the repeat sequence.

# BLAST\_EVALUE (M) E-value of the the BLASTN alignment between the contig and the repeat sequence.

# RC\_OF\_HSP\_EXTRACT (N) Orientation of the genome extract (HSP with 3' and 5' flanks) that was used for calculating the hybrid structure.

# HSP\_EXTRACT\_ID (O) ID of the genomic extract (not very important).

# RH\_MFE (P) Free energy (calculated by RNAHybrid) of the repeat-tracrRNA hybrid.

# HYBRID\_SEQS (Q) The repeat and the anti-repeat part of the tracrRNA (calculated by RH) separated by "&".

# HYBRID\_STRUCT (R) Base-pairings of the hybrid (repeat and anti-repeat, calculated by RH) in dot-bracket format.

# TAIL\_SEQ (S) The sequence of the tail part (i.e. after the anti-repeat) of the tracrRNA (This column is important).

# TAIL\_STRUCT (T) The folding of tracrRNA tail (predicted by RNAfold, RF).

# REPEAT\_TRACR (U) Repeat and the tracrRNA separated by "&".

# RH\_RF\_STRUCT (V) Base-pairings of the hybrid (by RH) in dot-bracket format, and the folding of tracrRNA tail (by RF) joined together.

# RCF\_STRUCT (W) Base-pairings of the hybrid and the folding of tracrRNA tail predicted simultaneously by RNAcofold (RCF).

# TRACR\_RNA\_SEQ (X) TracrRNA sequence (This column is important).

# TRACR\_START (Y) The start-coordinate of the tracrRNA.

# TRACR\_END (Z) The end-coordinate of the tracrRNA.

# TRACR\_ORI (AA) Whether or not the tracrRNA is on the reverse strand (1 if yes, 0 otherwise).

# N\_CONTIG (AB) Number of contigs in the GBFF file (a structured genomic DNA file).

# CONTIG\_LEN (AC) Length of the contig.

# CAS9\_START (AD) The start-coordinate of Cas9 (from PSI-BLAST).

# CAS9\_END (AE) The end-coordinate of Cas9.

# CAS9\_ID (AF) The SWISS-PROT ID of the predicted Cas9.

# CAS9\_ORI (AG) Direction of the Cas9 gene (1 if it is on the forward strand , 0 otherwise).

# CAS9\_PRED\_TYPE (AH) Description (part of) of the predicted Cas9 (not very important).

# DIST\_TRACR\_CAS (AI) The spacing between Cas9 and tracrRNA (unable to calculate if there a Cas9 gene is not predicted, or the Cas9 gene and the tracrRNA are on different contigs, in this case it will be 10^8).

# DIST\_TRACR\_RR (AJ) The spacing between tracrRNA and repeat array (unable to calculate if the repeat array and the tracrRNA are on different contigs, in this case it will be 10^8).

# DIST\_RR\_CAS (AK) The spacing between Cas9 and repeat array (unable to calculate if there a Cas9 gene is not predicted, or the Cas9 gene and the repeat array are on different contigs, in this case it will be 10^8).

# DIR\_SCORE (AL) Score of the choice of repeat-array direction. (0 if we unnecessarily take the reverse-complement of the repeat (and this is when the repeat always starts with GYY, or not start with GYY, regardless of taking the reverse-complement, in this case CRISPRDetect's direction prediction is reasonable but we rebutted it). 1 otherwise).

# TS\_SCORE (AM) Score = 1 if the tail of the tracrRNA starts with AR, 0 otherwise.

# TAIL\_FOLD\_SUMMARY (AN) A summary of the tail folding. (L = linker, H = hairpin, C = complex fold (hairpin within hairpin), number = number of nucleotides).

# NUM\_H (AO) Number of hairpins.

# MIN\_H (AP) Length of the smallest hairpin (0 if none).

# MAX\_H (AQ) Length of the largest hairpin (0 if none).

# FIRST\_H (AR) Length of the first hairpin (0 if none, and also, the first folding in the tail may be functionally important according to experimental studies).

# NUM\_C (AS) Number of complex fold (hairpin within hairpin).

# MIN\_C (AT) Length of the smallest complex fold (0 if none).

# MAX\_C (AU) Length of the largest complex fold (0 if none).

# FIRST\_C (AV) Length of the first complex fold (0 if none).

# NUM\_L (AW) Number of complex linker.

# MIN\_L (AX) Length of the smallest linker (0 if none).

# MAX\_L (AY) Length of the largest linker (0 if none).

# FIRST\_L (AZ) Length of the first linker (0 if none).

# TETRA\_T\_END (BA) Whether or not the tracrRNA ends with a tetra-T (1 if yes, 0 if no).

# XRAY\_FOLD (BB) Whether or not the tracrRNA tail contains 2-3 8-40nt hairpins without complex fold, and ends with tetra-T (this is based on the X-ray crystallographic structure of the S pyogenes and S aureus tracrRNAs, 1 if yes, 0 if no).

# PR\_POS (BC) SVM (support vector machine) true-positive probability.

# PR\_NEG (BC) SVM true-negative probability.

# Columns 57-71 (BE-BS) These are PATSCAN results, for conserved sequences that are known (not considered in this study).

# OVERLAPPING (BT) Whether or not the tail of the tracrRNA overlaps with a gene (1 if yes, 0 if no).

# OL\_GENE (BU) The gene locus ID, if the tail of the tracrRNA overlaps with a gene.

# CHY (BV) Whether or not the tracrRNA is reported by Chylinski et al. or has experimental evidence (1 if yes, 0 if no).

# THREE\_WAY\_JUNCT (BW) The three-way junction part of the tracrRNA, including the last 15 bases of the anti-repeat, and the first 15 bases of the tracrRNA tail.

# N\_PRED (BX) Number of predictions from the genome.

# S4 PDF file (Genome context of tracRNAs.pdf) shows the genomic context of tracRNAs. Genomic sequences within 5000 bases of a predicted tracrRNA gene (Red) were extracted and annotated using PROKKA. CRISPR-arrays (green), hypothetical and known genes including *cas9* (blue) were displayed.

S5 PDF file (Taxtree\_all\_tracrRNA.pdf) shows clusters of tracrRNA tails, repeat, anti-repeat and the full-length tracrRNA across genera. The relationships between the genera shown are based on NCBI’s taxonomic database.

S6 Additional validation statistics (S6 Additional validation statistics.docx) of the tracrRNA prediction pipeline. The false discovery rate (FDR) and the sensitivity were calculated at the level of genome. To calculate the FDR, we counted the number of genomes with at least one tracrRNA prediction within 3501-17500 bases of a *cas9* gene, divided by the number of genomes with at least one tracrRNA prediction within 17500 bases of a *cas9* gene. To calculate the sensitivity, we counted the number of genomes with a *cas9* gene (>2000 bases) and a tracrRNA, divided by the number of genomes with a *cas9* gene (>2000 bases).

# S6 Tail clusters. Stockholm files of RNAAlifold consensus structure of CLUSTALW-aligned UCLUST clusters (identity cutoff 0.65) of tracrRNA tails.

# S7 Covariation models of tracrRNA tails clusters in one text file (Tail\_Clusters\_cm.txt).

# S8 Repeat clusters. CLUSTALW alignments of CRISPR repeat clusters generated by UCLUST (identity cutoff 0.65).