Identification and evolution of Cas9 tracrRNAs

Shane K. Dooley[1†]*, Erica K. Baken[2], Walter Moss[3], Adina Howe[4] & Joshua K. Young[5]*

1 Department of Agricultural and Biosystems Engineering, Iowa State University, 605 Bissell Rd. Ames, IA 50011, USA, +1 (208) 881-1090, dooley.shanek@gmail.com
2 Department of Science, Chatham University, Buhl Hall, Chapel Hill Rd. Pittsburgh, PA 15232, +1 (763) 242-5479, erica.baken@gmail.com
3 Department of Biochemistry, Biophysics and Molecular Biology, Iowa State University, 2437 Pammel Drive Ames, IA 50011, USA, +1 (515) 294-6214, wmoss@iastate.edu
4 Department of Agricultural and Biosystems Engineering, Iowa State University, 605 Bissell Rd. Ames, IA 50011, USA, +1 (515) 294-0176, adina@iastate.edu
5 Department of Molecular Engineering, Corteva Agriscience™, 8305 NW62nd Ave Johnston, IA 50131, USA, +1 (515) 330-5986, josh.young@corteva.com
† First author
* To whom correspondence should be addressed.

Keywords: Cas9, tracrRNA, identification, classification, co-evolution, phylogeny

Abstract

Cas9 trans-activating CRISPR RNAs (tracrRNAs) form distinct structures essential for target recognition and cleavage and dictate exchangeability between orthologous proteins. As non-coding RNAs that are often apart from the CRISPR array, their identification can be arduous. In this paper, a new bioinformatic method for the detection of Cas9 tracrRNAs is presented. The approach utilizes a co-variance model (CM) based on both sequence homology and predicted secondary structure to locate tracrRNAs. This method predicts a tracrRNA for 98% of CRISPR-Cas9 systems identified by us. The identified tracrRNAs exhibit wide variation in sequence identity, however, CM analyses allow 94.7% to be categorized into just 10 related groups. Finally, association between Cas9 amino acid sequence-based phylogeny and tracrRNA secondary structure is evaluated, revealing strong evidence that secondary structure is evolutionarily conserved among Cas9 lineages. Altogether, our findings provide insight into Cas9 tracrRNA evolution and efforts to characterize the tracrRNA of new Cas9 systems.

Introduction

CRISPR (clustered regularly interspaced palindromic repeats) RNA (crRNA) and CRISPR-associated (Cas) proteins cooperate to defend prokaryotic organisms against invading RNA and DNA.[1,2] The Cas9 protein from type II CRISPR-Cas systems are guided to cleave double-strand (ds)DNA targets using two non-coding (nc)RNAs, a crRNA and a tracrRNA (trans-activating crRNA).[3,4] The crRNA contains a sequence, termed the spacer, that directly base pairs with the dsDNA target site in the vicinity of a protospacer adjacent motif (PAM).[5–7] The tracrRNA base pairs with the crRNA and is recognized and

bound by Cas9 resulting in the formation of a dual guide RNA (gRNA) ribonucleoprotein (RNP) complex.[8,9]

In recent years, due to its RNA-based programmability, CRISPR-Cas9 has been widely adopted as a genome editing tool for a variety of different genomes including those from eukaryotic organisms.[8,10,11] For these applications, the repair of a Cas9 induced double-strand break (DSB) has been harnessed to correct disease-causing mutations, introduce beneficial modifications (e.g. plant grain yield), and construct new biosynthetic pathways.[12–16] To further simplify its use, the dual gRNA has been engineered into a single gRNA (sgRNA) by linking the crRNA and tracrRNA.[8] Modifications to the Cas9 protein itself have also been made. By fusing new proteins domains to it and impairing its nuclease activity, it has been used as a robust RNA-guided DNA-binding platform. These applications include gene transcriptional activation and repression, epigenomic alteration, and precise DNA target deamination and modification.[17–24]

In prokaryotes, thousands of Cas9s have been identified computationally.[25–28] In contrast, the gRNA solution for orthologous Cas9s may not be easily recognizable. This is mainly due to large variation in tracrRNA location, size and sequence identity.[27,29] Consequently, the identification of tracrRNAs represents a bottleneck for the characterization of new Cas9 proteins and their development as genome editing tools. To address this limitation, several approaches have been developed. These include computational methods that locate tracrRNAs by using the CRISPR repeat sequence to search for the sequence in the tracrRNA that has homology to and base pairs with the crRNA (the anti-repeat). This is followed by a search for a Rho-independent-like termination signal in the vicinity of the anti-repeat. Other approaches reliant on the sequencing of the small ncRNAs transcribed from the CRISPR-Cas9 locus have also been used.[27,30]

Here, an algorithm was devised that combines previous methods with searches based on sequence homology and secondary structure co-variant models (CMs) to identify Cas9 tracrRNAs. Using this approach, a tracrRNA was located for greater than 98% of all CRISPR-Cas9 containing assemblies identifiable by us. Moreover, based on both sequence and structural homology, 94.7% of identified tracrRNAs could be unified into only 10 groups. Finally, Bayesian and non-parametric approaches quantifying phylogenetic signal revealed a strong evolutionary association between the Cas9 phylogeny and the predicted secondary structure of the tracrRNA, confirming previous observations that tracrRNA structures are a main determinant of Cas9-gRNA compatibility.[29,31]

Materials and Methods

All custom code, scripts, parsers, python objects, and Jupyter Notebooks can be found on the primary author's GitHub repository (https://github.com/skDooley/TRACR_RNA).

Detection of CRISPR-Cas9 systems

Bacterial and archaeal assemblies were downloaded from PATRIC2, NCBI GenBank and RefSeq (last downloaded on May 05, 2020). CRISPR arrays were identified using MinCED v0.3.2 and PilerCR v1.06 with relaxed parameter settings (3 or more crRNAs, repeat lengths between 16 and 64 base pairs, and max spacer lengths of 64 base pairs).[32,33]

Next, a Hidden-Markov model (HMM) was generated from 83 previously described diverse Cas9 proteins using HMMER 3.2.1.[27,34] The HMM was then used to search for Cas9-like proteins encoded in assemblies containing a CRISPR array. Protein sequences for each assembly were generated by translating open reading frames (ORFs) using Biopython to generate and filter ORFs for sequences between 673 and 2100 amino acids (Figure 1).[35] Next, using BLAST, assemblies duplicated in our collection were removed.[36]

The remaining assemblies and their Cas9 homologs' were further examined for the presence of RuvC and HNH cleavage domains that define a Cas9 nuclease.[8,9] This was initially accomplished through the visual inspection of protein alignments performed using MUSCLE between 83 diverse Cas9s described earlier for the key catalytic amino acids defining RuvCI, II and III subdomains and the HNH domain.[27,37] Next, the identified regions were extracted and used to generate domain specific HMMs using HMMER 3.2.1. Each putative Cas9 protein from our collection was then scanned with the cleavage domain-specific HMMs. Proteins missing either domain or that had subdomains that were positional outliers were removed. Outlier determination was made by assessing the position of the RuvC I subdomain near the N-terminus and then comparing the relative distance of all other cleavage domains. Anything outside of three standard deviations (distribution of all the search results for the RuvC I subdomain) was removed except for the RuvC III subdomain, where proteins with more than four standard deviations from the mean distance were removed.

For phylogenetic signal analysis, the translated sequences were clustered at 90% sequence homology using CD-HIT v4.7.[38] Representative sequences within each cluster were then selected and subsampled for calculating Cas9 and tracrRNA phylogenetic signal.

Identification of Cas9 tracrRNAs

Step1: Search for anti-repeat signatures

The region of the tracrRNA capable of base pairing with the crRNA, the anti-repeat, was identified in Cas9-containing assemblies by searching for sequences with homology to the CRISPR repeat (using BLAST 2.7.3) that were in a region distinct from the CRISPR array (Figure 1: Step 1). While all assemblies had CRISPR arrays, neither of the two programs (PilerCR or MinCED) accurately detected all of the CRISPR repeats. To correct this and significantly reduce false positives, the coordinates of putative anti-repeats in the locus were referenced and used to identify locations that were at least one repeat-spacer unit length away from the CRISPR array.

Step 2: Detect Rho-independent termination signals

Rho-independent-like termination signals (RTS) were detected using ERPIN v5.5 (parameters -add 1 4 1 and -cutoff 100%) and an RTS database (Figure 1, Step 2).[39] For this, the up- and down-stream regions adjacent to the anti-repeat were scanned for the presence of an RTS. Initially, each anti-repeat candidate with its respective RTS was considered a viable tracrRNA candidate. Additionally, if an anti-repeat had a termination signal on both sides, the pair were considered as potential tracrRNAs. Next, all tracrRNA candidates were conservatively filtered by removing sequences whose combined length was greater than 300 base-pairs. This cutoff-value was based on the longest characterized Cas9 tracrRNA length plus a generous buffer.[25]

Step3: Clustering tracrRNA candidates

Putative tracrRNAs were next clustered at 95% sequence identity with a 90% sequence coverage cutoff using cd-hit-est v.4.7.[38] Sequence clusters that did not map back to at least five different assemblies were removed from further analysis unless they contained a putative tracrRNA from an assembly with only a single putative candidate (Figure 1, Step 3). The resulting sequences and their respective clusters then formed the basis for structural predictions.

Steps 4 and 5. tracrRNA structural predictions and searches for orthologous sequences

To generate a consensus secondary structure for each sequence-based tracrRNA cluster, sequences from each cluster were first aligned using MAFFT (--maxiterate 1000 --globalpair) and then fed into RNAalifold 2.4.5 (Figure 1, Step 4).[40] The resulting consensus folds and sequences were then used as co-variance models within INFERNAL 1.1.2 to find RNA orthologs within the Cas9 associated DNA assemblies identified earlier (Figure 1, Step 5).[41] All results of the CM search were then filtered to remove any hits whose corresponding nucleotide sequence had less than 55% pairing with either the consensus repeat or the reverse complement of the repeat.

Step 6. Analysis of CM overlap

Following tracrRNA identification, a final analysis was performed to examine the overlap between CMs. For this, CMs from steps 4 and 5 were used with INFERNAL 1.1.2 to identify similarities between each putative tracrRNA sequence cataloged in Steps 1-5. Results were next visualized by creating an undirected graph. In the graph, CMs were represented as vertices and a line was added between the two vertices if the CMs identified the same putative tracrRNA sequence. Connecting line widths were scaled by the percentage of shared sequences (percent similarity=(# of shared sequences)/(min(# found with model 1,# found with model 2))). Each network was then pruned for lines separating weakly connected vertices in order to isolate highly similar clusters. For phylogenetic analyses, all clusters not associated with the top 10 most common structures were removed to make statistical calculations computationally feasible.

Calculating phylogenetic signal

To estimate the degree to which tracrRNA secondary structure associations are evolutionarily conserved among Cas9 lineages, phylogenetic signal was quantified using both Bayesian and non-parametric approaches. For the Bayesian approach, ancestral states of tracrRNA secondary structures along the Cas9 phylogeny were estimated using maximum likelihood under an All-Rates-Differ model in the R package diversitree.[42] Achieving convergence with the full dataset was unattainable due to the computational complexity of estimating transition rates with more than 10 discrete tracrRNA states. Thus, the original dataset was pruned to include only the 10 most common tracrRNA secondary structures (as described above). Subsequently, this dataset was subsampled to represent 25% of the pruned data (512 lineages) while preserving the 62 verified tracrRNA sequences.[43] With the ancestral state estimates, phylogenetic delta was calculated using a time-continuous discrete-trait Markov chain models (2 chains, 100,000 iterations each, thinned every 10 iterations, 100 iterations deleted as burn-in, see Borges et al. 2019 for more details). Values of phylogenetic delta above 1 indicate a close correspondence of the trait with the phylogeny, with increasing values representing increasing correspondence (i.e., strong phylogenetic signal), whereas values near 0 indicate weak phylogenetic signal. The results presented below were generated from a subsampling procedure that successfully converged. The Cas9 lineages involved in this calculation can be found in the supplemental materials (Supplementary Table S1). To ensure the results were not biased by subsampling, ten iterations were performed, and the resulting delta values for each round of subsampling can be found in the supplemental materials (Supplementary Table S2).

The second approach for quantifying phylogenetic signal was a modified two-block partial least squares test.[44] This procedure utilized the pruned dataset described above prior to the resampling procedures (2050 lineages included) and quantified the correlation coefficient of the Cas9 phylogeny (converted to a

phylogenetic covariance matrix) with the trait matrix. Multivariate effect size and significance were calculated using residual randomization via permutation procedures (1000 iterations, R package geomorph).[45,46]

Results

Identification of type II CRISPR-Cas9 systems

41,999 putative type II CRISPR-Cas9 systems from over 1 million microbial nucleotide sequences were identified (Supplementary Table S3). Cas9 length ranged from 700 to 1,800 amino acids and exhibited a bimodal length distribution centered around 1,100 and 1,400 amino acids (Supplementary Figure S1). To calculate the phylogenetic relationship between tracrRNA and Cas9, 2,724 diverse and representative systems were also selected (Supplementary Table S3) and subsampled (Methods section and Supplementary Table S1). Additionally, as a control for our methods, 79 type II CRISPR-Cas9 systems with an experimentally established tracrRNA were also included in our analysis (Supplementary Table S4).[43] Of these, a Cas9 encoding ORF was detected for 73 using our methods.

Co-variant detection of Cas9 tracrRNAs

The detection of Cas9 tracrRNAs was automated using a multi-step approach that combines both homology and structural searches (Figure 1). First, taking advantage of previous methods, the identification of the tracrRNA anti-repeat and Rho-independent termination-like signal were automated similar to those described in Chyou *et al*., 2019 (Figure 1, Steps 1 and 2).[27,47–49] Next, based on functional associations between gRNA secondary structure and orthogonality, we reasoned that conserved tracrRNA structural features could be used to complement homology-dependent methods in the identification of a tracrRNA.[31] To accomplish this, sequences of tracrRNAs predicted in Steps 1 and 2 (Figure 1) were first aligned and clustered based on sequence similarity (Figure 1, Step 3). Next, sequences within each cluster were used to predict a consensus secondary structure (Figure 1, Step 4). CMs were next generated from each cluster based on sequence and structural homology and used to search for related tracrRNAs (Figure 1, Step 5). Finally, to examine the relationship between tracrRNAs in our collection, a last clustering step based on CM similarity was applied (Figure 1, Step 6). For some systems, multiple solutions were observed within the CRISPR-Cas9 locus after Step 6 (Figure 1) (Supplementary Figure S2A-E). In these cases, additional filtering was applied to permit the selection of a single tracrRNA. First, the tracrRNA that was closest to the *cas9* gene was selected. If two or more tracrRNAs in the region closest to the *cas9* gene had overlapping locations, the tracrRNA with the most stable secondary structure (based of minimum free energy calculations of the predicted folds) was chosen.

Next, our pipeline was used to predict tracrRNA solutions for the 41,999 CRISPR-Cas9 systems identified earlier. To establish false positive and negative rates of our approach, its ability to accurately predicted the tracrRNA from a curated set of 74 experimentally proven Cas9 tracrRNAs was also evaluated.[43] For this, the loci containing the curated tracrRNAs were identified and flagged in our collection. Altogether, our algorithm predicted a tracrRNA for 98% (41741 out of 41999) of the Cas9 systems searched (Supplementary Table S3). For the curated set of tracrRNAs proven to support Cas9 functionality, our approach correctly identified 90.4% (66 out of 73) resulting in false positive and negative rates of 6.8% (5 out of 73) and 2.7% (2 out of 73), respectively. Of the five systems where a different tracrRNA was identified, four (Cco, Kki, Lsp2 and Nsa) were predicted to have a tracrRNA that was transcribed in the opposite direction from the anti-repeat than described earlier. For the fifth system, the tracrRNA was predicted to be in a different location (Ghy3) (Supplementary Figure S2A-E).[43]

Sequence and structural homology of Cas9 gRNAs

Based on sequence and structural overlap, 94.7% (39527 out of 41741) of identified tracrRNAs could be categorized into 10 clusters (Figure 2). 1,388 of the 2,214 remaining tracrRNAs were classified into 31 additional CM-based similarity groups and 826 of the remainders represented as singletons in the dataset (Supplementary Table S3). The majority of previously characterized tracrRNAs could be found in the 10 most abundant clusters (Figure 2, Clusters 1-7 and 10).

To visualize tracrRNA structural features in the context of the dual gRNA used by Cas9, a sgRNA was generated by linking the 3' end of the full-length CRISPR repeat with a self-folding tetraloop (5'-GAAA-3') to the 5' end of the anti-repeat in the tracrRNA as described previously.[50] This was done once for each of the top 10 clusters using the most abundant tracrRNA sequence and respective CRISPR repeat. As observed previously, most sgRNA structures comprised varying degrees of complementation between the repeat and anti-repeat followed by two or more hairpin-like structures in the tracrRNA (Figure 2).[29,31,43] Likewise, a repeat:anti-repeat mismatch resulting in a bulge was detected in some but not all instances (Figure 2). In most cases, the nexus fold, a functionally important and conserved hairpin structure hypothesized to orient the spacer away from the rest of the dual gRNA, was detected almost immediately (within 2 or 3 nts) after the repeat:anti-repeat duplex (Figure 2, Clusters 1, 2, 4-6, 7 and 9)[25,31]. For clusters 3, 8 and 10, it was located approximately 9 nts after the repeat:anti-repeat (Figure 2). The nexus-like fold itself varied in length from 10-80 nts with an average length of 24 nts and ranged from simple 3 nt stem loop structures (Figure 2, Clusters 1, 4 and 6) to more complex structures with additional bulges and stems (Figure 2, Clusters 3 and 9).

Cas9 and tracrRNA evolutionary association

Cas9 phylogeny and predicted gRNA secondary structures have been linked to exchangeability between orthologous Cas9s.[29,31] This suggests tight evolutionary association between Cas9 and its gRNA structural features. To further test this observation, we examined the phylogenetic association between tracrRNA secondary structure and Cas9 protein. For this, two statistical methods, Bayesian estimation of the phylogenetic delta statistic and a non-parametric modified two-block partial least squares model, were used to evaluate the phylogenetic relationship between the 10 primary tracrRNA secondary structures (encompassing 83.3% of all representative tracrRNAs identified) and our representative collection of Cas9 proteins. First, a diverse and representative collection of Cas9s were subsampled and a phylogenetic tree was constructed. Next, tracrRNA structures were mapped to it (Figure 3). Ancestral states were then estimated, from which the delta statistic was calculated. In these scenarios, a significant phylogenetic signal was detected using both approaches (delta = 244.107 and r-PLS = .912, effect size = 28.952, p = 1e-04) as can be visualized by the strong clustering of tracrRNA secondary structures across Cas9 phylogeny (Figure 3). Rare exceptions to this were observed as the occurrence of the same tracrRNA structure in distantly related orthologs (Figure 3).

Discussion

We provide a framework for the global identification of CRISPR-Cas9 tracrRNAs. Our method builds upon previous approaches [27,47–49] that have sought to identify the key components that define a tracrRNA, the anti-repeat, and 3' hairpin-like secondary structures, and adds to them by utilizing CMs to identify sequence and structural homologs (Figure 1, Steps 1-5). In total, we predicted a tracrRNA solution from 98% of the identified Cas9 systems. In comparison with a diverse collection of experimentally determined tracrRNAs, we also showed that our approach in most cases (66 out of 73 (90.4%)) could accurately identify a Cas9 tracrRNA.[43] In the five instances where a different tracrRNA was detected (Cco, Ghy3, Kki, Lsp2 and Nsa), it is possible that a second tracrRNA may have evolved in the CRISPR-Cas9 locus as described previously.[25] This is supported in part by the identification of alternative tracrRNAs in these loci that exhibit CM homology to tracrRNAs known to support Cas9 functionality (Figure 2 and Supplementary Figures S2A-E). Additionally, the location of the alternate tracrRNA within the CRISPR-Cas9 locus is consistent with other characterized systems. These locations include regions near the end of the CRISPR array or directly adjacent to the *cas9* gene (Supplementary Figures S2A-E).

In examining the sequence and structural overlap of the identified tracrRNAs using CMs (Figure 1, Step 6), we found that they could be classified mainly into 10 groups (Figure 2). Interestingly, when observing the distribution for previously determined tracrRNAs, it seems that our structural classifications also correlated with Cas9-gRNA compatibility (Figure 2).[29,43,51] Although, in some cases, as observed in

cluster 2, Cas9 and gRNAs (Spy and Tde) previously shown to be incompatible were grouped together (Figure 2).[51] This finding indicates that features in the repeat:anti-repeat duplex may also confer non-compatibility or that some clusters represent a continuum of related tracrRNA structures that diverge into non-compatible ones. Indeed, the latter point seems more likely for Spy and Tde provided the related, yet almost distinct groupings observed within cluster 2 (Figure 2). Altogether, our findings suggest the number of non-cross reactive gRNA groupings may be extended from 7 to 10 or more.[31,43,51] Furthermore, this finding expands the number of potential Cas9s available for orthogonal genome editing approaches or applications that require RNA-guided multiplexing.[51,52]

Both approaches for calculating phylogenetic signal showed that the tracrRNA structure is an evolutionarily conserved trait among Cas9 lineages. This matches previous observations that Cas9-gRNA exchangeability is associated with gRNA secondary structure and Cas9 phylogeny.[29,31] Interestingly, exceptions to this were noted in our analysis. In those instances, distantly related Cas9 orthologs were associated with the same tracrRNA structural classification. This observation may provide evidence of more recent evolutionary events resulting from the resetting of the tracrRNA or recombination between different CRISPR-Cas9 systems.[25]

Conclusion

Using CMs based on both sequence homology and predicted structure, an informatic approach, enabling the identification of Cas9 tracrRNAs, was developed. This method permitted the global identification of more than 41K tracrRNAs and the development of sgRNA solutions for nearly all CRISPR-Cas9 systems detected by us. Structural predictions revealed strong homology among tracrRNA secondary structure that tightly correlated with the Cas9 phylogeny allowing the majority of Cas9 tracrRNAs to be classified into just 10 groups. Altogether, the results presented here will aid in the characterization and development of new Cas9s as genome editing tools and maybe extended to other CRISPR systems that utilize a tracrRNA.[53–56]

Acknowledgements

We thank Dr. Dean Adams (Iowa State University) for his advice and expertise on how to calculate phylogenetic signal and for his quick replies to bugs in the phylogenetic analysis. Additionally, thank you to Kevin Hayes for supporting the early stages of this analysis and Corteva for funding the initial stages of this research. We thank the ISU legal team and Lynne Mumm for her work in facilitating the collaboration agreements between Corteva and Iowa State. Finally, thank you to Dr. James Reece and the Office of the Vice President for Research at Iowa State University for funding me (SKD).

Authorship Confirmation Statement

SKD, EKB, WM, AH and JKY designed research; SKD performed research and SKD, EKB and JKY analyzed data. SKD, EKB, AH, and JKY wrote the paper. All authors read and approved the final manuscript and it has not been published, in press, or submitted elsewhere.

Author Disclosure Statement

SKD, EKB, WM, and AH have no competing financial interests. JKY is an employee of Corteva Agriscience.

References

1. Mojica, F. J. M., Díez-Villaseñor, C., García-Martínez, J. & Soria, E. Intervening sequences of regularly spaced prokaryotic repeats derive from foreign genetic elements. *J. Mol. Evol.* (2005). doi:10.1007/s00239-004-0046-3

2. Pourcel, C., Salvignol, G. & Vergnaud, G. CRISPR elements in Yersinia pestis acquire new repeats by preferential uptake of bacteriophage DNA, and provide additional tools for evolutionary studies. *Microbiology* (2005). doi:10.1099/mic.0.27437-0

3. Deltcheva, E. *et al.* CRISPR RNA maturation by trans-encoded small RNA and host factor RNase III Europe PMC Funders Group. *Nature* (2011). doi:10.1038/nature09886

4. Garneau, J. E. *et al.* The CRISPR/cas bacterial immune system cleaves bacteriophage and plasmid DNA. *Nature* (2010). doi:10.1038/nature09523

5. Deveau, H. *et al.* Phage response to CRISPR-encoded resistance in Streptococcus thermophilus. *J. Bacteriol.* (2008). doi:10.1128/JB.01412-07

6. Horvath, P. *et al.* Diversity, activity, and evolution of CRISPR loci in Streptococcus thermophilus. *J. Bacteriol.* (2008). doi:10.1128/JB.01415-07

7. Mojica, F. J. M., Díez-Villaseñor, C., García-Martínez, J. & Almendros, C. Short motif sequences determine the targets of the prokaryotic CRISPR defence system. *Microbiology* (2009). doi:10.1099/mic.0.023960-0

8. Jinek, M. *et al.* A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science (80-.).* (2012). doi:10.1126/science.1225829

9. Gasiunas, G., Barrangou, R., Horvath, P. & Siksnys, V. Cas9-crRNA ribonucleoprotein complex mediates specific DNA cleavage for adaptive immunity in bacteria. *Proc. Natl. Acad. Sci.* (2012). doi:10.1073/pnas.1208507109

10. Cong, L. *et al.* Multiplex genome engineering using CRISPR/Cas systems. *Science (80-.).* (2013). doi:10.1126/science.1231143

11. Mali, P. *et al.* RNA-guided human genome engineering via Cas9. *Science (80-.).* (2013). doi:10.1126/science.1232033

12. Schwank, G. *et al.* Functional repair of CFTR by CRISPR/Cas9 in intestinal stem cell organoids of cystic fibrosis patients. *Cell Stem Cell* **13**, 653–658 (2013).

13. Wu, Y. *et al.* Correction of a Genetic Disease in Mouse via Use of CRISPR-Cas9. *Cell Stem Cell* **13**, 659–662 (2013).

14. Shi, J. *et al.* ARGOS8 variants generated by CRISPR-Cas9 improve maize grain yield under field drought stress conditions. *Plant Biotechnol. J.* **15**, 207–216 (2017).

15. Wang, Z. *et al.* CRISPR-Cas9 HDR system enhances AQP1 gene expression. *Oncotarget* **8**, 111683–111696 (2017).

16. Jakočinas, T. *et al.* Multiplex metabolic pathway engineering using CRISPR/Cas9 in Saccharomyces cerevisiae. *Metab. Eng.* (2015). doi:10.1016/j.ymben.2015.01.008

17. Gilbert, L. A. *et al.* XCRISPR-mediated modular RNA-guided regulation of transcription in eukaryotes. *Cell* (2013). doi:10.1016/j.cell.2013.06.044

18. Mali, P. *et al.* CAS9 transcriptional activators for target specificity screening and paired nickases for cooperative genome engineering. *Nat. Biotechnol.* (2013). doi:10.1038/nbt.2675

19. Perez-Pinera, P. *et al.* RNA-guided gene activation by CRISPR-Cas9-based transcription factors. *Nat. Methods* (2013). doi:10.1038/nmeth.2600

20. Hilton, I. B. *et al.* Epigenome editing by a CRISPR-Cas9-based acetyltransferase activates genes from promoters and enhancers. *Nat. Biotechnol.* (2015). doi:10.1038/nbt.3199

21. Gaudelli, N. M. *et al.* Programmable base editing of A•T to G•C in genomic DNA without DNA cleavage. *Nature* **551**, 464–471 (2017).

22. Morgan, S. L. *et al.* Manipulation of nuclear architecture through CRISPR-mediated chromosomal looping. *Nat. Commun.* (2017). doi:10.1038/ncomms15993

23. Zhou, Y. *et al.* Painting a specific chromosome with CRISPR/Cas9 for live-cell imaging. *Cell Research* (2017). doi:10.1038/cr.2017.9

24. Anzalone, A. V. *et al.* Search-and-replace genome editing without double-strand breaks or donor DNA. *Nature* (2019). doi:10.1038/s41586-019-1711-4

25. Faure, G. *et al.* Comparative genomics and evolution of trans-activating RNAs in Class 2 CRISPR-Cas systems. *RNA Biology* (2018). doi:10.1080/15476286.2018.1493331

26. Mohanraju, P. *et al.* Diverse evolutionary roots and mechanistic variations of the CRISPR-Cas systems. *Science* (2016). doi:10.1126/science.aad5147

27. Chylinski, K., Le Rhun, A. & Charpentier, E. The tracrRNA and Cas9 families of type II CRISPR-Cas immunity systems. *RNA Biol.* (2013). doi:10.4161/rna.24321

28. Shmakov, S. *et al.* Diversity and evolution of class 2 CRISPR-Cas systems. *Nat. Rev. Microbiol.* (2017). doi:10.1038/nrmicro.2016.184

29. Fonfara, I. *et al.* Phylogeny of Cas9 determines functional exchangeability of dual-RNA and Cas9 among orthologous type II CRISPR-Cas systems. *Nucleic Acids Res.* (2014). doi:10.1093/nar/gkt1074

30. Deltcheva, E. *et al.* CRISPR RNA maturation by trans-encoded small RNA and host factor RNase III. *Nature* (2011). doi:10.1038/nature09886

31. Briner, A. E. *et al.* Guide RNA functional modules direct Cas9 activity and orthogonality. *Mol. Cell* (2014). doi:10.1016/j.molcel.2014.09.019

32. Bland, C. *et al.* CRISPR Recognition Tool (CRT): A tool for automatic detection of clustered regularly interspaced palindromic repeats. *BMC Bioinformatics* (2007). doi:10.1186/1471-2105-8-209

33. Edgar, R. C. PILER-CR: Fast and accurate identification of CRISPR repeats. *BMC Bioinformatics* (2007). doi:10.1186/1471-2105-8-18

34. Eddy, S. R. Accelerated profile HMM searches. *PLoS Comput. Biol.* (2011). doi:10.1371/journal.pcbi.1002195

35. Cock, P. J. A. *et al.* Biopython: Freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* (2009). doi:10.1093/bioinformatics/btp163

36. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* (1990). doi:10.1016/S0022-2836(05)80360-2

37. Edgar, R. C. MUSCLE: A multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* (2004). doi:10.1186/1471-2105-5-113

38. Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. CD-HIT: Accelerated for clustering the next-generation sequencing data. *Bioinformatics* (2012). doi:10.1093/bioinformatics/bts565

39. Gautheret, D. & Lambert, A. Direct RNA motif definition and identification from multiple sequence alignments using secondary structure profiles. *J. Mol. Biol.* (2001). doi:10.1006/jmbi.2001.5102

40. Lorenz, R. *et al.* ViennaRNA Package 2.0. *Algorithms Mol. Biol.* (2011). doi:10.1186/1748-7188-6-26

41. Nawrocki, E. P. & Eddy, S. R. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics* (2013). doi:10.1093/bioinformatics/btt509

42. Fitzjohn, R. G. Diversitree: Comparative phylogenetic analyses of diversification in R. *Methods Ecol. Evol.* (2012). doi:10.1111/j.2041-210X.2012.00234.x

43. Gasiunas, G. *et al.* Biochemically diverse CRISPR-Cas9 orthologs. *bioRxiv* 2020.04.29.066654 (2020). doi:10.1101/2020.04.29.066654

44. Rohlf, F. J. & Corti, M. Use of two-block partial least-squares to study covariation in shape. *Syst. Biol.* (2000). doi:10.1080/106351500750049806

45. Collyer, M. L. & Adams, D. C. RRPP: An r package for fitting linear models to high-dimensional data using residual randomization. *Methods Ecol. Evol.* (2018). doi:10.1111/2041-210X.13029

46. Adams, D. C. & Collyer, M. L. Phylogenetic ANOVA: Group-clade aggregation, biological challenges, and a refined permutation procedure. *Evolution (N. Y).* (2018). doi:10.1111/evo.13492

47. Karvelis, T. *et al.* Rapid characterization of CRISPR-Cas9 protospacer adjacent motif sequence elements. *Genome Biol.* (2015). doi:10.1186/s13059-015-0818-7

48. Karvelis, T. *et al.* PAM recognition by miniature CRISPR-Cas14 triggers programmable double-stranded DNA cleavage. *bioRxiv* 654897 (2019). doi:10.1101/654897

49. Chyou, T. yuan & Brown, C. M. Prediction and diversity of tracrRNAs from type II CRISPR-Cas systems. *RNA Biol.* (2019). doi:10.1080/15476286.2018.1498281

50. Jinek, M. *et al.* A Programmable Dual-RNA – Guided. *Science* (2012).

doi:10.1126/science.1225829

51. Esvelt, K. M. *et al.* Orthogonal Cas9 proteins for RNA-guided gene regulation and editing. *Nat. Methods* (2013). doi:10.1038/nmeth.2681

52. Morgan, S. L. *et al.* Manipulation of nuclear architecture through CRISPR-mediated chromosomal looping. *Nat. Commun.* **8**, 15993 (2017).

53. Zetsche, B. *et al.* Cpf1 Is a Single RNA-Guided Endonuclease of a Class 2 CRISPR-Cas System. *Cell* (2015). doi:10.1016/j.cell.2015.09.038

54. Burstein, D. *et al.* New CRISPR-Cas systems from uncultivated microbes. *Nature* (2017). doi:10.1038/nature21059

55. Harrington, L. B. *et al.* Programmed DNA destruction by miniature CRISPR-Cas14 enzymes. *Science* **362**, 839–842 (2018).

56. Yan, W. X. *et al.* Functionally diverse type V CRISPR-Cas systems. *Science (80-.).* (2019). doi:10.1126/science.aav7271

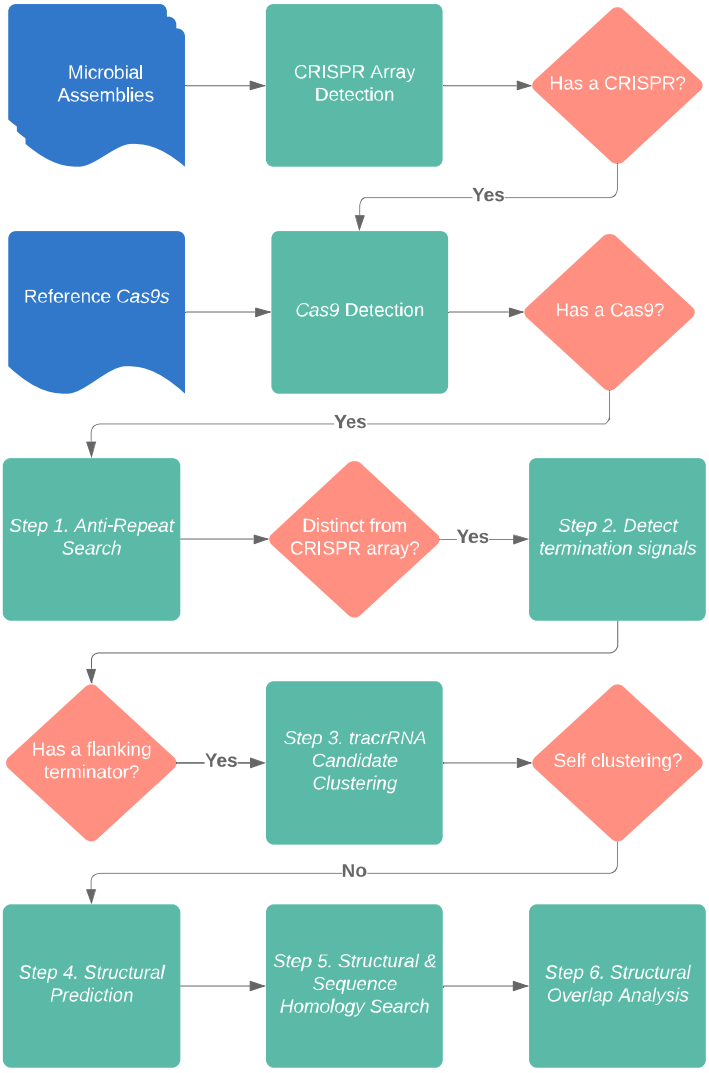

Figure 1. Cas9 tracrRNA detection pipeline.
Flowchart of the informatic steps and key decisions points used to predict Cas9 tracrRNAs. Cas9-containing CRISPR systems are first identified in microbial DNA assemblies. Assemblies with a CRISPR-Cas9 loci are then searched in 6 steps to predict a tracrRNA. Inputs are shown in blue, informatic activities indicated in green and key decision points highlighted in orange.

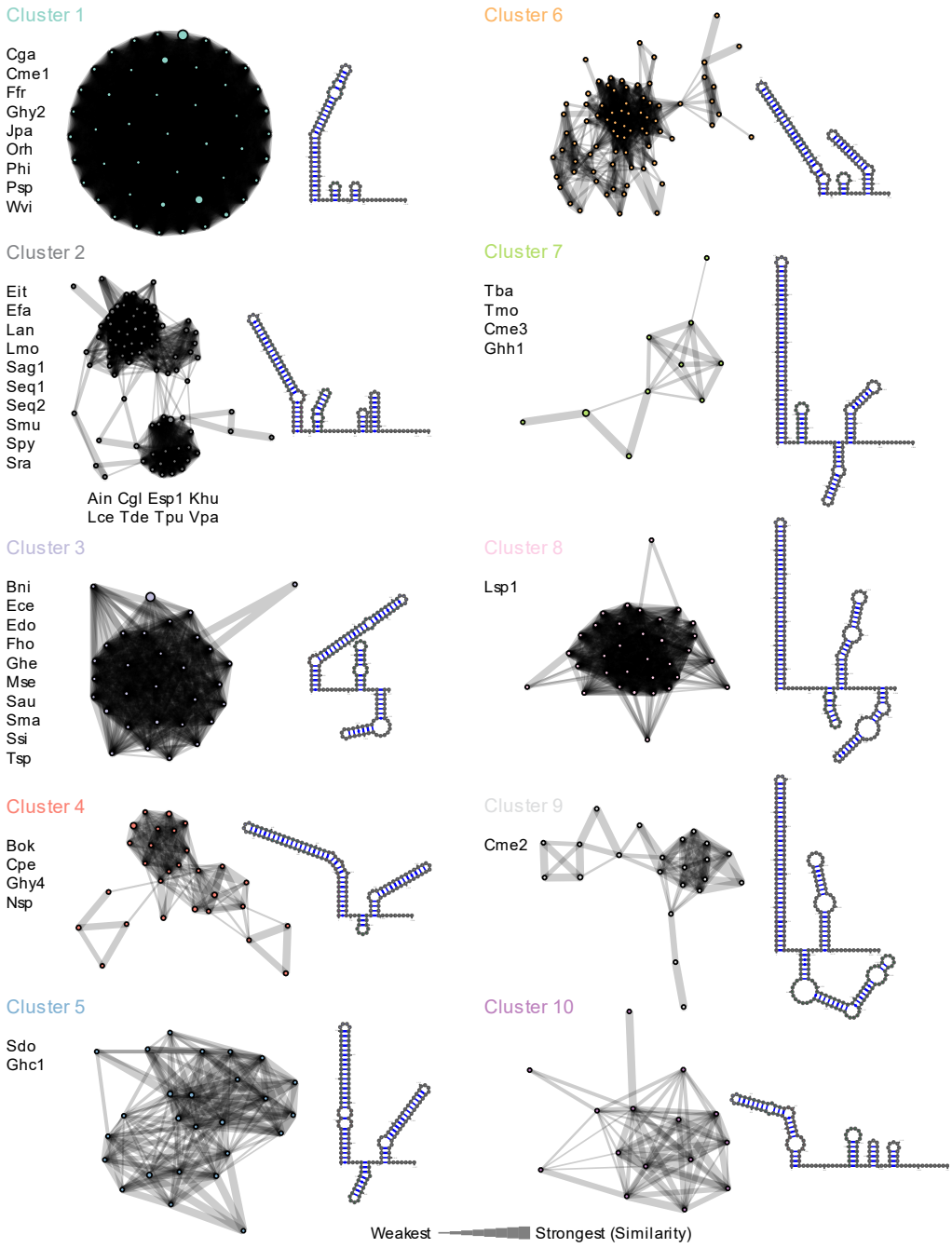

Figure 2. Top 10 co-variant models and clustering of Cas9 tracrRNAs

Top 10 Cas9 tracrRNA clusters based on similarity between sequence and predicted secondary structure co-variant models (CMs). Circles represent a CM and are colored according to the designated cluster. The width of the connecting lines indicates the percentage of similarity or relatedness among CMs. Previously characterized tracrRNAs associated with each cluster are indicated. Single guide RNA (sgRNA) solutions for the most abundant tracrRNA sequence are also shown immediately adjacent to each cluster. The lower stem, bulge, upper stem and nexus regions are colored purple, orange, teal, blue and green, respectively.

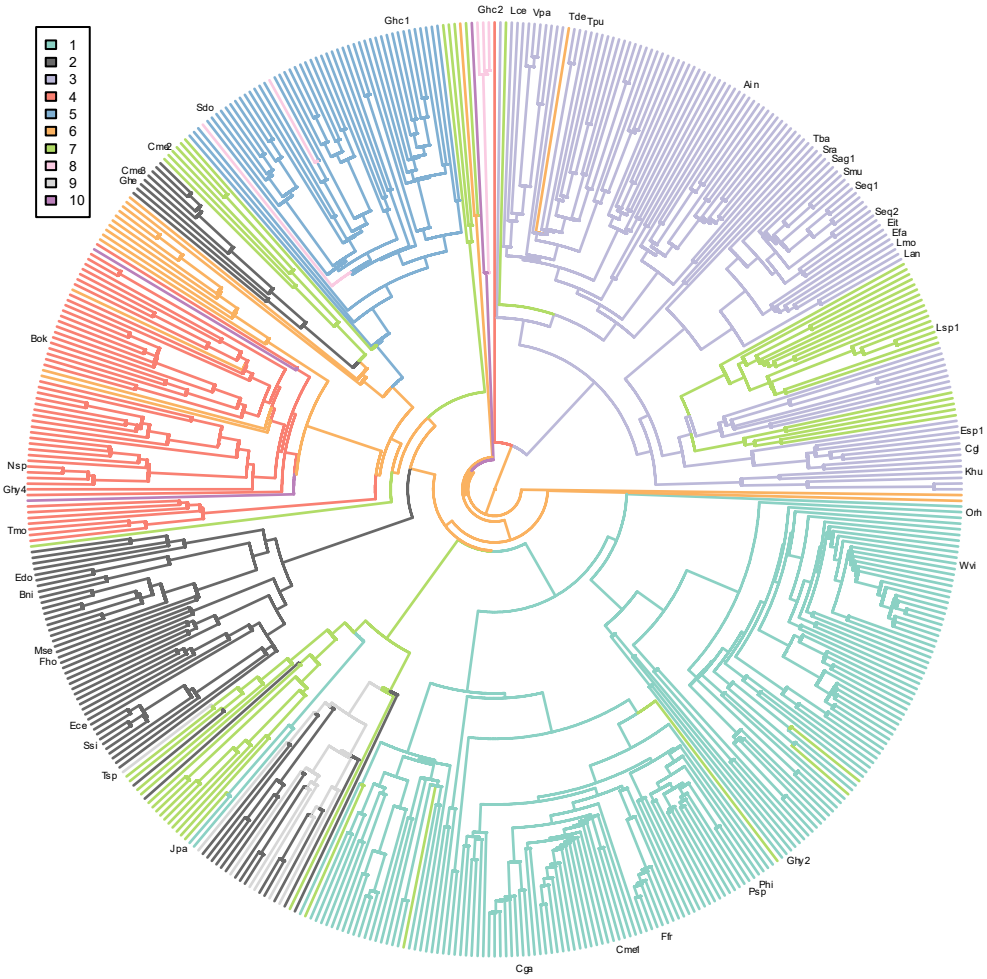

Figure 3. Cas9 phylogeny and tracrRNA secondary structure
Predicted tracrRNA secondary structures associated with Cas9 phylogeny. Each color represents tracrRNA secondary structure associated with Clusters 1-10 (Figure 2). Cas9 proteins characterized previously are indicated (Supplementary Table S4).