CRISPRtracr supplementary materials

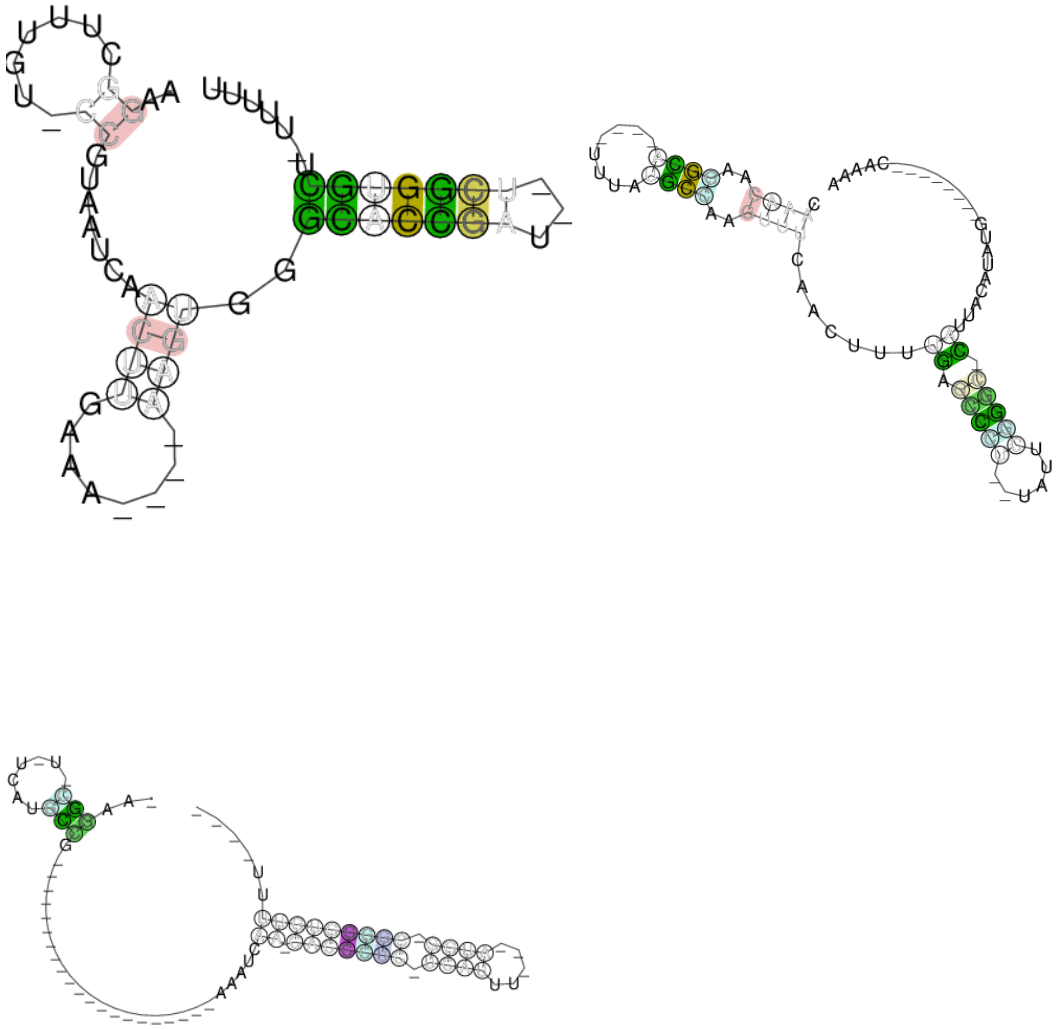

Figure S1. Consensus structure of the tail sequences for 3 clusters. The representation is taken directly from the GraphClust2 pipeline.

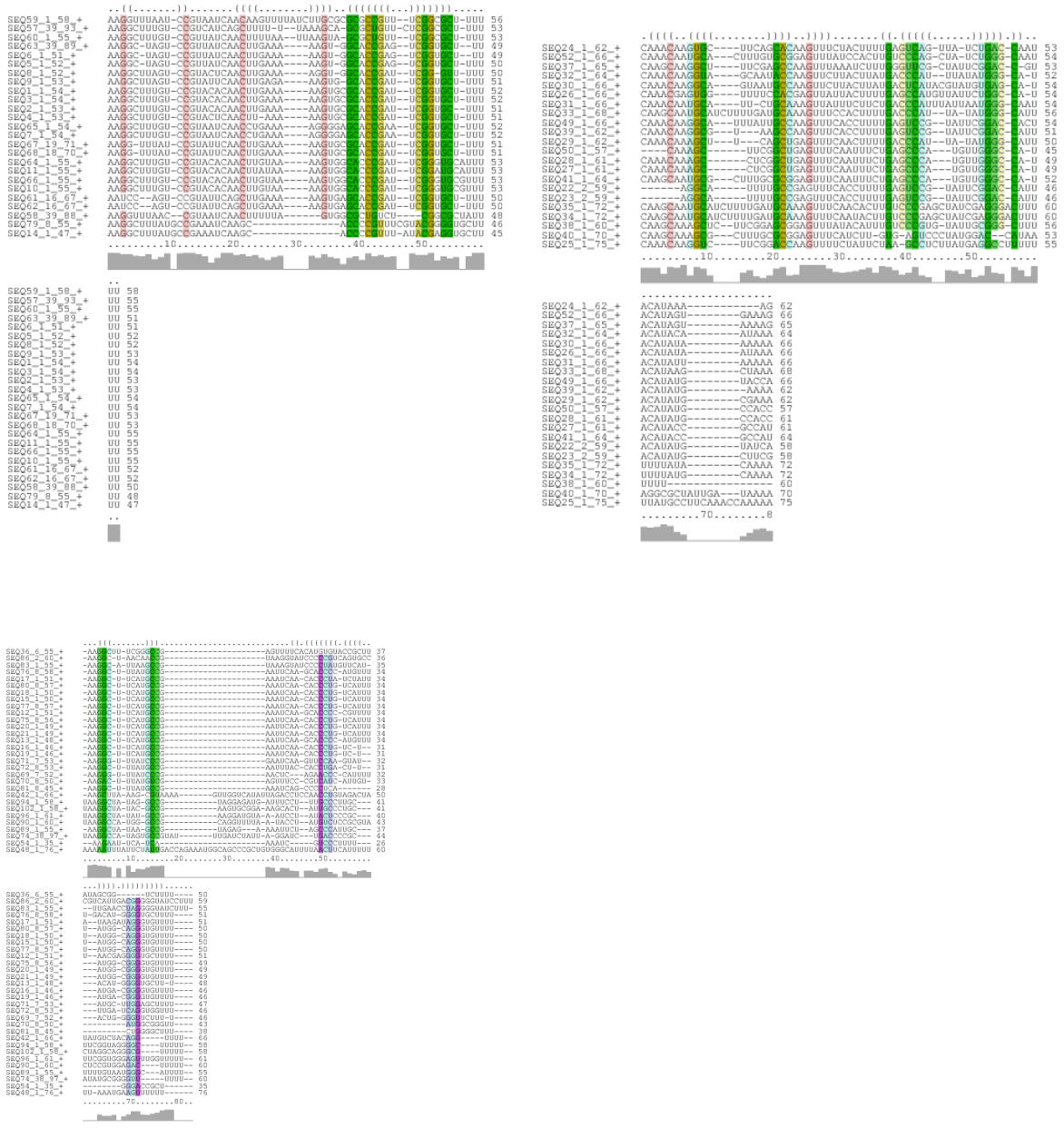

Figure S2. Sequence constitution for the 3 clusters obtained by GraphClust2.

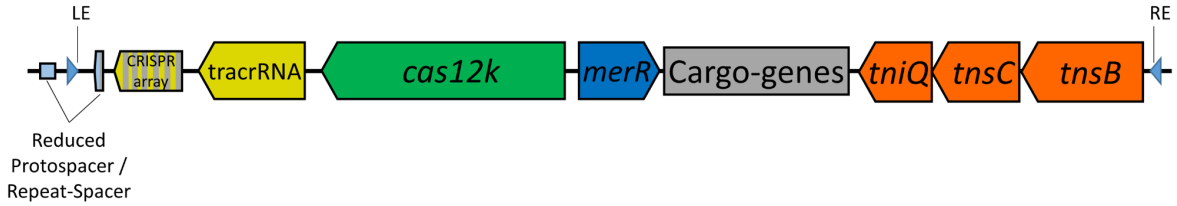

Figure S3. Gene map of common CAST system
The CAST transposon is displayed from its left insertion element (LE) to its right insertion element (RE). The genes are colored according to their function (green: *cas12k*, orange: Transposase, blue: *merR*-regulator, yellow: RNA transcribing regions). The CRISPR array is depicted with its Repeats (grey) and Spacers (yellow) separated from the reduced repeat-spacer pair (~12 nt repeat, ~17 nt spacer), corresponding to a protospacer upstream of the LE-element (light blue).

```
Query:       1  [L=189]
Hit scores:
 rank     E-value  score  bias  modelname                                    start    end   mdl trunc   gc  description
 ----   --------- ------ -----  ---------------------------------------- ------ ------   --- ----- ----  -----------
  (1) !   3.7e-41  131.2   0.0  dataset_ac9e1c46-3bdb-438c-b4e4-d87c82a0bf98     88    189 +  cm    no 0.60  -
  (2) !   3.6e-36  111.9   0.0  dataset_8c0eca81-d482-4680-8a18-52bf60a8c927     90    189 +  cm    no 0.59  -
  (3) !   1.6e-18   65.9   0.0  dataset_bb909aed-2e23-46c9-b48a-cfc70a24cc0b    106    189 +  cm    no 0.58  -
  (4) !   2.9e-18   58.0   0.0  dataset_ea4e8ca6-a768-42a6-9697-12da98e5493e     92    189 +  cm    no 0.59  -
  (5) !   3.2e-18   57.6   0.0  dataset_8f025680-be32-4e39-800f-241dd6189933     90    189 +  cm    no 0.59  -
  (6) !   8.8e-18   55.5   0.0  dataset_d82cbec9-f11a-44f7-90f8-eb3ad81f52af      1    102 +  cm    no 0.50  -
  (7) !   1.7e-17   59.2   0.0  dataset_c223a4d9-f85f-481c-99c9-d740a2614d96      1    102 +  cm    no 0.50  -
  (8) !   3.9e-14   44.7   0.0  dataset_d5f82872-fcbb-48d9-bd73-534e21290062      1     83 +  cm    no 0.47  -
  (9) !   3.3e-13   44.1   0.0  dataset_2b49d5c0-6516-4a11-9d43-bdd362a64f01     52    129 +  cm    no 0.60  -
 (10) !   5.4e-11   34.8   0.0  dataset_d88cd498-b5c3-46f6-bc9d-cd1b2694ca46     52    127 +  cm    no 0.59  -
 (11) !     8e-10   34.4   0.0  dataset_0d97656a-f089-4cea-b254-1037e8105f1a      1    102 +  cm    no 0.50  -
 (12) !   2.6e-09   33.9   0.0  dataset_337bf8ea-be72-4180-b7cd-c98b90660b72     52    129 +  cm    no 0.60  -
 (13) !   4.3e-09   28.9   0.0  dataset_56eb54d6-41b5-4287-951f-8c211a177cfe     52     84 +  cm    no 0.52  -
 (14) !    0.0001   17.9   0.0  dataset_5e0a073e-da8c-46f7-8ac4-f8b97f52eed1    102    127 +  cm    no 0.65  -
 (15) !    0.0098   11.7   0.0  dataset_80e8b885-fe88-4767-a3d9-7e67ae2f2891     61    128 +  cm    no 0.59  -
 ------ inclusion threshold ------
 (16) ?       3.3    1.0   0.0  dataset_0d97656a-f089-4cea-b254-1037e8105f1a    143    146 +  cm    3' 0.50  -
 (17) ?       3.4    0.9   0.0  dataset_d5f82872-fcbb-48d9-bd73-534e21290062     90    102 +  cm    no 0.69  -
```

Figure S4. Infernal cmscan hits of the type V validation dataset tracrRNA candidates with the pretrained model.

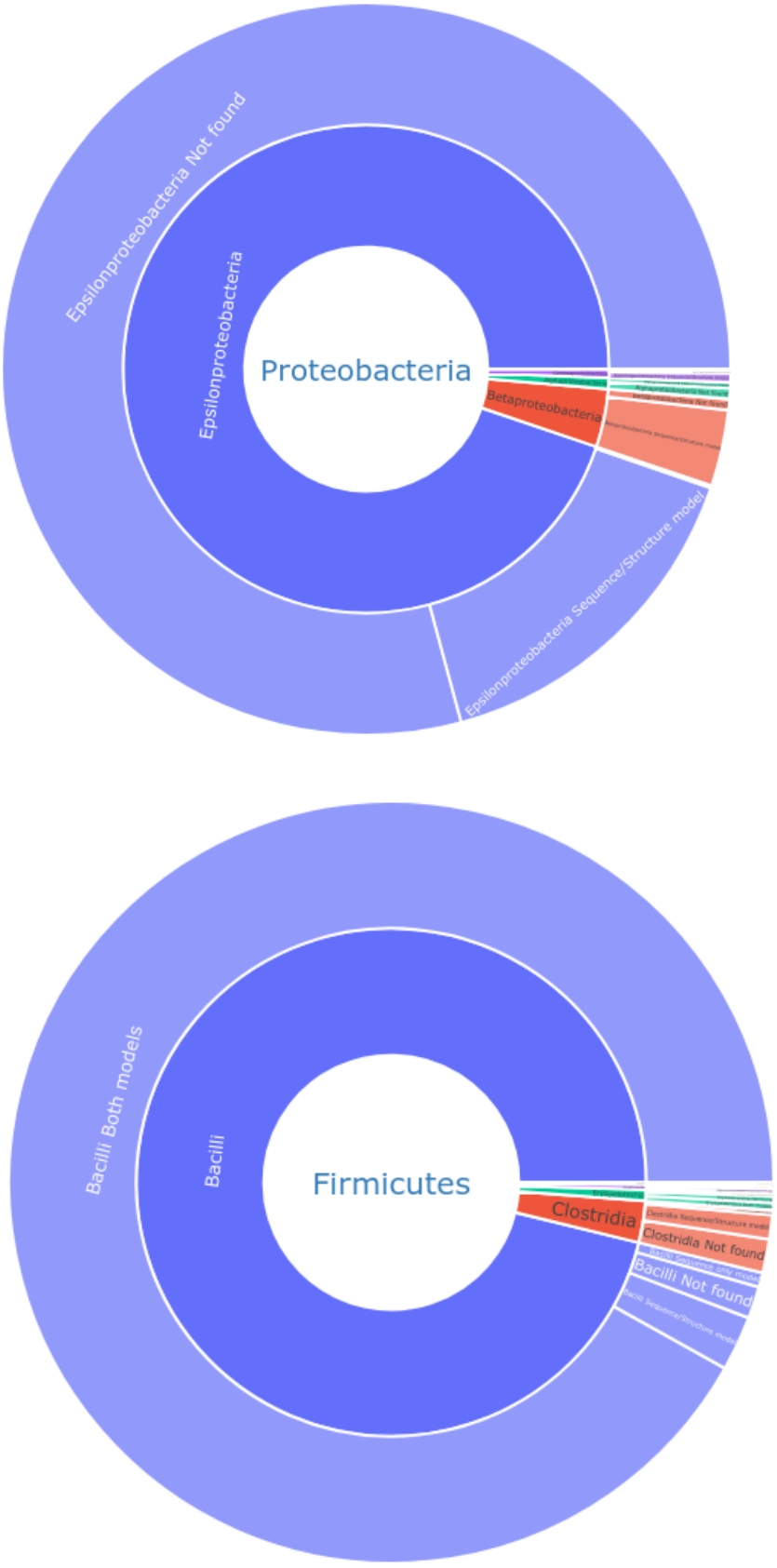

Figure S5. Distribution of the found candidates of sequence only and sequence/structure models in the Proteobacteria and Firmicutes phylums.

Evidence factor	Default importance weigh for candidate ranking (not a threshold value)
CRISPR array certainty score	0.5
Anti-repeat BLAST similarity	0.5
Anti-repeat BLAST coverage	0.5
BLAST similarity * coverage	0.5
IntaRNA score	0.6
CM Model hit score	0.9
Terminator score	0.9
Consistency orientation	0.1
Consistency anti-repeat/tail	0.1
Consistency tail/terminator	0.1

Table S1 The default values for evidence component importance. The user might assign the values between 0 and 1 to adjust the output according to their priority.

More data is available on https://github.com/BackofenLab/CRISPRtracrRNA