Supplementary text 1: Distribution of Spacer Hits: Contributions of Microbial Diversity and Diversity of CRISPR-Cas systems

The entire set of identified CRISPR arrays, all the matching protospaces and, separately, the subset of virus protospacers were broken down by the CRISPR-Cas system subtype and by the microbial phyla (Table X1-X3). For all 88 phylum/subtype categories with at least 10 matches, the fraction of spacers with matches and the fraction of virus matches among all matches were calculated. Both of these fractions were analyzed as follows.

The calculated fraction $f_{i,j}$ for the spacers in phylum *i* and belonging to the category *j* was modeled as the product of a phylum-specific factor p_i and a subtype-specific factor s_j. The deviation of the observed and the predicted fractions was assumed to be log-normally distributed with the expectation of 1. Formally,

$$\log f_{i,j} = \log p_i + \log s_j + e_{i,j}$$

where $e_{i,j} \sim \text{Norm}(0,\sigma)$.

Optimization of vectors ***p*** and ***s*** with respect to Σe^2 provides both the phylum- and subtype-specific contribution factors and the residuals $e_{i,j}$ that allow one to estimate the quality of the fit between the model and the observation. Setting one of the vectors to an arbitrary positive value (e.g. 1) and optimizing the other provides a model with phylum- or subtype-only contributions; the null model replaces both vectors with mean($\log f_{i,j}$). The sums of squared residuals can be compared using the Fisher F-test.

Contribution of phylum- and subtype-specific factors to the fraction of spacers with matches

	phylum- and subtype	subtype-	subtype-	None
No. of data points	88	88	88	88
No. of parameters	39	21	18	0
No. of degrees of freedom	47	65	68	85
Σe^2	4.32	7.17	13.02	18.30

Contribution of phylum- and subtype-specific factors to the fraction of spacer matches to virus sequences

	phylum- and subtype	**subtype-**	**subtype-**	**None**
No. of data points	88	88	88	88

No. of parameters	39	21	18	0
No. of degrees of freedom	47	65	68	85
Σe^2	0.24	0.40	0.66	0.79

For both the overall fraction of matches and the fraction of viral matches, both phylum- and subtype-specific factors provide significant and independent contributions to the reduction of residuals (Fisher F-test p-value <0.05 except p=0.07 for the contribution of subtype-specific factors alone to the overall fraction of matches). Taken together, the two factors explain 70-75% of the original variance in the observed fraction.

Number of spacers in CRISPR arrays

Phylum	I	I-A	I-B	I-C	I-D	I-E	I-F	I-U	II-A	II-B	II-C	III	III-A	III-B	III-C	III-D	IV-A	V-A	V-B	VI-A	VI-B	VI-C	V-U	?
Acidithiobacillia	0	0	0	17	0	0	0	0	0	0	6	0	0	0	0	0	58	0	0	0	0	0	3	0
Acidobacteriia	0	0	0	21	0	0	0	33	0	0	0	0	0	0	14	0	0	0	0	0	0	0	0	6
Actinobacteria	1224	59	1613	3660	0	26325	0	3163	455	0	1504	188	608	150	0	899	368	0	0	0	0	0	19	14640
Alphaproteobacteria	14	7	0	2092	52	2931	43	333	0	0	893	0	75	99	0	100	24	0	8	5	0	0	0	1459
Aquificae	0	0	410	0	0	0	0	0	0	0	0	10	25	0	6	5	0	0	0	0	0	0	0	492
Archaeoglobi	4	131	91	0	0	0	0	0	0	0	0	9	99	0	0	0	0	0	0	0	0	0	0	174
Armatimonadetes	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	151
Bacteria candidate phyla	0	0	60	0	0	0	0	23	0	0	23	0	64	40	0	5	2	34	0	0	0	0	0	256
Bacteroidetes/Chlorobi group	118	138	5529	1496	0	486	108	48	28	0	4767	205	584	619	61	498	0	226	0	25	682	0	0	2993
Betaproteobacteria	84	15	2	2774	120	3345	2676	58	0	24	1447	67	140	284	16	60	97	0	0	0	0	0	0	2804
Blastocatellia	0	0	0	0	0	0	0	50	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100
Caldisericia	0	0	22	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Caldithrix	0	0	30	0	0	0	0	0	0	0	0	0	36	0	0	12	0	0	0	0	0	0	0	11
Candidatus Korarchaeota	0	117	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Candidatus Marinimicrobia	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	85
Candidatus Tectomicrobia	0	0	0	0	0	11	0	0	0	0	0	0	0	0	0	4	0	0	0	0	0	0	0	19
Chlamydiae	0	0	0	0	0	23	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Chloroflexi	16	4	2248	1014	684	190	0	3	0	0	0	279	605	56	4	151	127	0	0	0	0	0	0	1142
Chrysiogenetes	0	0	0	0	0	17	38	0	0	0	0	0	54	0	0	0	0	0	0	0	0	0	0	0
Cloacimonetes	0	0	180	0	0	0	0	0	0	0	6	72	0	0	0	24	0	0	0	0	0	0	0	19
Crenarchaeota	941	4794	0	0	388	0	0	0	0	0	0	265	437	1047	87	1273	16	0	0	0	0	0	0	1964
Cyanobacteria/ Melainabacteria group	60	119	2806	256	4923	482	0	292	0	0	0	1014	132	3404	55	1265	183	0	0	0	0	0	256	5048
Deferribacteres	0	0	114	0	0	0	69	0	0	0	10	0	55	31	0	0	0	0	0	0	0	0	0	22
Deinococcus-Thermus	80	141	438	444	0	1423	0	33	0	0	0	88	177	229	0	4	49	0	0	0	0	0	14	937
delta/epsilon subdivisions	172	0	2320	2862	254	2256	487	1115	0	46	1256	155	301	321	125	132	12	6	6	0	0	0	3	1759
Dictyoglomia	0	0	90	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	110
Elusimicrobia	0	0	0	0	0	0	0	0	0	0	13	0	0	0	0	0	0	0	0	0	0	0	0	0
Endomicrobia	0	0	0	0	0	0	0	0	0	0	111	0	0	0	0	0	0	0	0	0	0	0	0	0
environmental samples	0	0	0	0	0	0	0	0	0	0	131	0	0	0	0	0	0	0	0	0	0	0	0	0
Fibrobacteres	0	0	0	119	0	19	31	0	0	0	45	0	0	24	0	21	0	0	0	0	0	0	0	156
Firmicutes	447	820	23433	15274	393	5328	378	178	12413	0	1493	1319	3448	2398	818	2117	194	94	131	140	0	3	20	22493
Fusobacteriia	4	0	1858	0	0	0	0	0	221	0	95	191	305	0	0	4	6	12	0	9	0	31	0	636
Gammaproteobacteria	1858	6	317	5533	285	14211	24247	1046	0	391	400	374	622	1331	0	501	330	61	0	0	0	0	0	15689
Gemmatimonadetes	0	0	0	5	0	11	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Hadesarchaea	0	0	52	0	0	0	0	0	0	0	0	0	21	0	0	0	0	0	0	0	0	0	0	0
Halobacteria	34	0	2974	0	497	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1616
Holophagae	0	0	0	26	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Latescibacteria	0	0	0	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3
Methanobacteria	0	0	1587	0	0	0	0	0	0	0	0	79	253	0	0	0	4	0	0	0	0	0	0	295
Methanococci	0	57	399	0	49	0	0	0	0	0	0	28	68	3	20	0	98	0	0	0	0	0	0	917
Methanomicrobia	233	81	3173	833	692	921	0	194	0	0	0	134	368	1048	426	834	0	0	0	0	0	0	0	1462
Methanopyri	0	0	0	0	0	0	0	0	0	0	0	9	0	0	0	0	0	0	0	0	0	0	0	17
nitrifying bacterium enrichment culture	0	0	0	0	0	114	0	94	0	0	0	0	0	136	0	0	0	0	0	0	0	0	0	0
Nitrospinae	0	0	0	0	0	0	0	0	0	0	21	0	0	0	0	0	0	0	0	0	0	0	0	0
Nitrospira	2	0	214	42	0	328	0	48	0	0	0	0	70	0	32	0	9	0	0	0	0	0	5	338
Planctomycetes	134	0	388	286	0	3	0	580	0	0	32	14	21	16	11	35	0	0	0	0	0	0	0	130
Solibacteres	0	0	0	64	0	0	0	0	0	0	22	0	0	0	0	0	0	0	0	0	0	0	0	30
Spirochaetia	170	10	882	802	53	942	46	0	71	0	245	0	33	20	0	6	0	51	0	0	0	0	0	1783
Synergistia	0	0	609	455	0	116	0	0	0	0	87	45	135	130	0	0	0	0	0	0	0	0	0	125
Tenericutes	0	0	0	0	0	0	0	0	755	0	111	0	2	0	0	0	0	0	0	0	0	0	0	128

Thaumarchaeota	0	21	252	0	0	0	0	0	0	0	0	22	0	0	0	41	0	0	0	0	0	0	0	225
Thermobaculum	0	0	0	0	0	18	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	35
Thermococci	59	416	495	0	0	0	0	0	0	0	0	0	153	117	0	0	0	0	0	0	0	0	0	1970
Thermodesulfobacteria	0	0	187	0	0	0	0	0	0	0	0	0	29	51	0	0	0	0	0	0	0	0	0	63
Thermoplasmata	13	0	117	112	194	32	0	84	0	0	0	0	63	0	0	0	0	108	0	0	0	0	0	547
Thermotogae	0	6	1226	0	43	0	0	0	0	0	0	72	212	368	242	50	0	0	0	0	0	0	0	1512
unclassified Acidobacteria	0	0	0	0	0	171	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	87
unclassified Archaea (miscellaneous)	0	0	0	0	60	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	8
unclassified Bacteria (miscellaneous)	3	0	631	123	408	17	0	0	24	0	237	3	33	196	0	289	0	0	0	0	0	0	0	850
unclassified Euryarchaeota	0	0	32	0	1	0	0	0	0	0	0	20	21	0	0	0	0	0	0	0	0	0	0	51
unclassified Proteobacteria	0	0	0	0	0	0	8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
uncultured archaeon	0	0	0	55	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Verrucomicrobia	0	0	0	201	0	63	0	119	0	0	67	0	0	123	0	15	0	0	23	0	0	0	0	80
Zetaproteobacteria	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5	0	0	0	0	0	0	25

Number of spacers with hits

Phylum	I	I-A	I-B	I-C	I-D	I-E	I-F	I-U	II-A	II-B	II-C	III	III-A	III-B	III-C	III-D	IV-A	V-A	V-B	VI-A	VI-B	VI-C	V-U	?
Acidithiobacillia	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0
Acidobacteriia	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
Actinobacteria	56	6	120	429	0	1716	0	287	20	0	145	3	11	1	0	11	63	0	0	0	0	0	0	746
Alphaproteobacteria	0	1	0	45	0	34	0	0	0	0	10	0	0	1	0	2	0	0	0	0	0	0	0	27
Aquificae	0	0	5	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	4
Armatimonadetes	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
Bacteria candidate phyla	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
Bacteroidetes/Chlorobi group	1	0	230	98	0	3	0	4	1	0	319	4	8	16	1	8	0	5	0	0	50	0	0	92
Betaproteobacteria	2	9	0	157	1	72	116	0	0	0	221	2	2	5	2	0	0	0	0	0	0	0	0	319
Candidatus Marinimicrobia	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2
Chloroflexi	0	0	1	7	0	16	0	0	0	0	0	0	2	0	0	2	0	0	0	0	0	0	0	2
Chrysiogenetes	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Cloacimonetes	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
Crenarchaeota	5	69	0	0	3	0	0	0	0	0	0	0	5	4	0	21	0	0	0	0	0	0	0	12
Cyanobacteria/ Melainabacteria group	0	0	25	1	35	3	0	1	0	0	0	4	2	16	0	4	0	0	0	0	0	0	1	34
Deinococcus-Thermus	0	5	6	9	0	38	0	4	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	21
delta/epsilon subdivisions	7	0	52	28	0	16	7	2	0	2	178	0	5	1	0	8	2	0	0	0	0	0	0	70
Elusimicrobia	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
environmental samples	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0
Fibrobacteres	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
Firmicutes	40	6	1380	978	11	338	7	0	2608	0	115	35	94	136	33	34	13	0	7	8	0	0	2	1798
Fusobacteriia	0	0	448	0	0	0	0	0	10	0	1	19	24	0	0	1	1	0	0	0	0	2	0	123
Gammaproteobacteria	286	0	0	529	0	1950	4736	12	0	7	46	0	13	45	0	15	65	0	0	0	0	0	0	2534
Halobacteria	0	0	47	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	26
Methanobacteria	0	0	37	0	0	0	0	0	0	0	0	1	6	0	0	0	0	0	0	0	0	0	0	3
Methanococci	0	1	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Methanomicrobia	9	0	36	14	9	8	0	1	0	0	0	3	2	31	4	14	0	0	0	0	0	0	0	10
nitrifying bacterium enrichment culture	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
Nitrospira	0	0	0	0	0	11	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2
Planctomycetes	1	0	1	9	0	0	0	1	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0
Spirochaetia	106	3	268	40	0	263	0	0	15	0	16	0	0	0	0	0	0	0	0	0	0	0	0	462
Synergistia	0	0	0	13	0	6	0	0	0	0	0	0	2	1	0	0	0	0	0	0	0	0	0	0
Tenericutes	0	0	0	0	0	0	0	0	25	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
Thaumarchaeota	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1
Thermococci	0	2	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	10
Thermodesulfobacteria	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
Thermoplasmata	0	0	0	0	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
Thermotogae	0	0	10	0	0	0	0	0	0	0	0	0	3	1	0	0	0	0	0	0	0	0	0	17
unclassified Bacteria (miscellaneous)	0	0	3	1	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0
unclassified Proteobacteria	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Verrucomicrobia	0	0	0	18	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	6

Number of spacers with virus hits

Phylum	I	I-A	I-B	I-C	I-D	I-E	I-F	I-U	II-A	II-B	II-C	III	III-A	III-B	III-C	III-D	IV-A	V-A	V-B	VI-A	VI-B	VI-C	V-U	?
Acidithiobacillia	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0
Acidobacteriia	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Actinobacteria	30	3	88	333	0	1317	0	226	18	0	107	1	5	1	0	7	58	0	0	0	0	0	0	552
Alphaproteobacteria	0	0	0	36	0	21	0	0	0	0	6	0	0	0	0	1	0	0	0	0	0	0	0	19
Aquificae	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4
Armatimonadetes	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
Bacteria candidate phyla	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
Bacteroidetes/Chlorobi group	1	0	209	70	0	2	0	4	1	0	244	3	6	7	1	2	0	5	0	0	36	0	0	66
Betaproteobacteria	1	2	0	130	0	65	92	0	0	0	156	2	0	3	0	0	0	0	0	0	0	0	0	178
Candidatus Marinimicrobia	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
Chloroflexi	0	0	1	5	0	13	0	0	0	0	0	0	1	0	0	2	0	0	0	0	0	0	0	1
Chrysiogenetes	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Cloacimonetes	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
Crenarchaeota	5	62	0	0	3	0	0	0	0	0	0	0	5	4	0	19	0	0	0	0	0	0	0	9
Cyanobacteria/ Melainabacteria group	0	0	20	1	27	3	0	1	0	0	0	3	1	10	0	4	0	0	0	0	0	0	1	23
Deinococcus-Thermus	0	4	3	4	0	31	0	4	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	17
delta/epsilon subdivisions	5	0	33	21	0	13	6	0	0	2	97	0	3	0	0	3	2	0	0	0	0	0	0	41
Elusimicrobia	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
environmental samples	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Fibrobacteres	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
Firmicutes	33	6	1271	870	6	295	7	0	2359	0	100	28	75	123	31	28	8	0	7	4	0	0	1	1628
Fusobacteriia	0	0	425	0	0	0	0	0	9	0	0	16	19	0	0	1	1	0	0	0	0	1	0	108
Gammaproteobacteria	254	0	0	465	0	1790	4378	12	0	2	32	0	6	42	0	13	35	0	0	0	0	0	0	2299
Halobacteria	0	0	27	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	13
Methanobacteria	0	0	24	0	0	0	0	0	0	0	0	1	5	0	0	0	0	0	0	0	0	0	0	3
Methanococci	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Methanomicrobia	9	0	34	13	8	6	0	0	0	0	0	2	2	30	4	14	0	0	0	0	0	0	0	7
nitrifying bacterium enrichment culture	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
Nitrospira	0	0	0	0	0	10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
Planctomycetes	0	0	1	9	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
Spirochaetia	68	2	189	28	0	206	0	0	10	0	13	0	0	0	0	0	0	0	0	0	0	0	0	338
Synergistia	0	0	0	8	0	5	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
Tenericutes	0	0	0	0	0	0	0	0	19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Thaumarchaeota	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Thermococci	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2
Thermodesulfobacteria	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Thermoplasmata	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Thermotogae	0	0	10	0	0	0	0	0	0	0	0	0	3	1	0	0	0	0	0	0	0	0	0	15
unclassified Bacteria (miscellaneous)	0	0	3	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
unclassified Proteobacteria	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Verrucomicrobia	0	0	0	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4

Supplementary text 2

Examples of intergenic hits into unannotated ORFs

Spacer 1:

>JODE01000009.1_62635_14_spacer_63457_32

GACCAGGCGTCCCCGCTGACGCTGATCGCGGT

Hit coordinates:
ContigID: LGDA01000250
Start: 45153
Stop: 45184

Intergenic sequence:
>gi|925387131|gb|LGDA01000250.1|:44655-45964 Streptomyces sp. WM6368 P399contig36.1, whole genome shotgun sequence
TAAGACATGCGGTGTGGCCGGGCACATGGATGCCGCGGCATCCGACAGGGTCCTGCGAGCTGCCGCCGTG
AGCACAGTGCGACCAGGCAGGCGGAGTCTGGCGGGGAACGGGAGTCGGCTGCGGCATACCCCGGAGGACG
ACTAAGTACAGCTGGATGGTTGCCGGCCACGCCCGTCGGTGGACCATCCGGACGGGCGGTCCTCGTCTGC
GGCGGCGTCGGCCAGGAGCAGGTGCGGGTCCTCCAGCCGGAACCGGCGCGTACCGCCCCCGGCTCCGGCT
CGTACCGGCGATACGGGGCGCTCTCGTAGACGACGGTGGTGTTCGGGCAGTCCTCCGCGACCCCGAGCAG
CGCTTGGCGCTCGGCGTCGTCGGCGGCAAGGCTCCAGCGCAACTTCGCCGCGGTCCACTGGGCGGCGTAC
TCGCAGTGGTAACCGCTGGCCGGCGGAATCCACTGCGCCGGATCCTTGTCGGCCTTCGACCGGTTCGACT
TCGCGGTCACCGCGATCAGCGTCAGCGGGGACGCCTGGTCGTTCGCGTACGCCTCACGCCGGACCGCGTC
CCACGCGAAGCTGCCCGAGTCGTGGACTTCGGCCAAAGGTACGAAATGGTCTACATCAAGGCCCGCGGCG
TCCGTCACAACGACGTCGCCGTAAGGGCTCAGCCGCGATCCACCCGACATCTTGCAGCCCGCAGCGACGA
CCGGGGCCTCGACGGCCTCGGACAGGATGACTTCCCTGCGGGTATCACAGCCGTCGGTCGCGTTCAGCCC
GCGGTTCCAGTGCTTGTACAGGTCCCACTTGTAACCCTCCCGCTGCTCATCAGCGACGGGGATCCGGTCG
ATGGCCTCGAACAGCGGCAGCGGCGCACGGACTCCGGGAGCCCCAGCGATACCGGACGAGGCGGGGACGC
CCGCCTGGGTGCCGGCCGCGTGCGCGGCGGGTGCGGTGGCGAGCAGGGGCAGGGCAGCGAGAGCGAGCGC
GGGAGGACCACGCTGCAGCAGGTTCTTGATCACGCAGTGGGTTGTAGCGGTGCCCCAGCGGGACAGGGCC
TGGGGCCTCAGCTCACCCGCACGGCCGCCGCGAATCACCGCAGAAACGATCAGCGGCTGGGCCGACCGGC
AGCCGGTCAACGCTCCTCAGCGCGAGATCATCTTTGGCCCGTGAGGCAACGGCTCGGTCCCGGGCAACCC
ACACGCATCATCGGGAGTACCAGCAGATTCCAGCGCACGGCACGTTGCAACATCATGGCGTCAGGTGCCT
CGAGTCAGCTGCACATGACGGTCTTTGCGACGTCTCCCAGACGAGGGCCG

Top 10 BLASTX hits: into HNH endonuclease
Fields: query id, subject id, subject ids, subject length, s. start, s. end, evalue, query seq, subject seq, q. start, q. end, score, subject sci names, subject title
10 hits found

gi|925387131|gb|LGDA01000250.1|:44655-45964 gi|919577089|ref|WP_052876068.1| gi|919577089|ref|WP_052876068.1| 236 21 235 2.97e-126
PAAHAAGTQAGVPASSGIAGAPGVRAPLPLFEAIDRIPVADEQREGYKWDLYKHWNRGLNATDGCDTRREVILSEAVEAPVVAAGCKMSGGSRLSPYGDVVVTDAAGLDVDHFVPLAEVHDSGSFAWDAVRREAYANDQASPLTLIAVTAKSNRSKADK
DPAQWIPPASGYHCEYAAQWTAAKLRWSLAADDAERQALLGVAEDCPNTTVVYESA
PTAHAEPVPATAHSTAGAPSAAGLRAPIPLFEAIDRIPVADEQREGYKRDLYKHWNRGLNATDGCDTRREVILSEAVEAPVVAAGCKMSGGSWLSPYDDVVVTDAAGLDVDHFVPLAEVHDSGGFAWDPARREAYANDQASPLTLIAVTAKSNRSKADK
DPAQWLPPARGYHCEYAAQWTATKLRWNLAADDAERQALLGVAEDCRNTTVVYESA941 297 958 Streptomyces sp. NRRL F-4335HNH endonuclease [Streptomyces sp. NRRL F-4335]
gi|925387131|gb|LGDA01000250.1|:44655-45964 gi|494479460|ref|WP_007268937.1| gi|494479460|ref|WP_007268937.1| 235 43 235 3.47e-110
VRAPLPLFEAIDRIPVADEQREGYKWDLYKHWNRGLNATDGCDTRREVILSEAVEAPVVAAGCKMSGGSRLSPYGDVVVTDAAGLDVDHFVPLAEVHDSGSFAWDAVRREAYANDQASPLTLIAVTAKSNRSKADKDPAQWIPPASGYHCEYAAQWTAA
KLRWSLAADDAERQALLGVAEDCPNTTVVYESAP
VRAPLPLFEAIDRITVAEEQREGYTRSLYKHWNRGLDATDGCDTRREVILSEALEAPQVTAGCKMSGGLWLSPYDDVTMTDAAGIDVDHFVPLAEVHDSGGYGWDAARREAYANDQASPDTLIAVTAKSNRSKADKDPAQWLPPAAGYHCQYAATWVGT
KLRWDLAADEAERQALLGLAEDCPDTTVVYEPAP 872 294 852 Streptomyces sp. C HNH endonuclease [Streptomyces sp. C]
gi|925387131|gb|LGDA01000250.1|:44655-45964 gi|739837920|ref|WP_037688867.1| gi|739837920|ref|WP_037688867.1| 234 1 234 9.32e-110
LQRGPPALALAALPLLATAPAAHAAGTQAGVPASSGIAGAPGVRAPLPLFEAIDRIPVADEQREGYKWDLYKHWNRGLNATDGCDTRREVILSEAVEAPVVAAGCKMSGGSRLSPYGDVVVTDAAGLDVDHFVPLAEVHDSGSFAWDAVRREAYANDQA
SPLTLIAVTAKSNRSKADKDPAQWIPPASGYHCEYAAQWTAAKLRWSLAADDAERQALLGVAEDCPNTTVVYESAP MFRGLPALALCALPILA-
APTAHSATAETAPLAATGVTGPAGVRAPLPLFEAIDRLPVAAEHRDGYKRDLYKHWNRGLNATDGCDTRKEVILAEAVTAPQVAAGCKLTSGSWLSAYDNVVVSDAARLDVDHFVPLAEVYDSERAPWSAARREAYANDQASPDTLIAVSAASNRSKADKDPAEWLPS
DDSYHCTYAASWVGTKLRWDLAVDESERHALLGLAEDCPATTVVYESAP 998 294 849 Streptomyces lavendulae HNH endonuclease [Streptomyces lavendulae]
gi|925387131|gb|LGDA01000250.1|:44655-45964 gi|1154964266|ref|WP_078869587.1| gi|1154964266|ref|WP_078869587.1| 226 1 225 1.18e-105
LQRGPPALALAALPLLATAPAAHAAGTQAGVPASSGIAGAPGVRAP--
LPLFEAIDRIPVADEQREGYKWDLYKHWNRGLNATDGCDTRREVILSEAVEAPVVAAGCKMSGGSRLSPYGDVVVTDAAGLDVDHFVPLAEVHDSGSFAWDAVRREAYANDQASPLTLIAVTAKSNRSKADKDPAQWIPPASGYHCEYAAQWTAAKLRWSLAADDAER
QALLGVAEDCPNTTVVYESA MLRGLPALVLAAVPLLATAPAAHA----------
GTALTPSSAQPVTLPLFDAVDQLPVADEQREGYQRSLYKHWTRGLNLTDGCDTRKEVILAEAVLAPTVTAGCRLSGGSWHSAYDDLTVTDAARLDVDHFVPLAEVHDSGGFAWDAKRREAYANDQDSPDTLIAVSAASNRSKADRDPAEWMPSDGSYHCTYTATWVAT
KLRWSLAADDTERQALLGLAEDCPATTVTYEPA 998 297 821 Streptomyces sp. NRRL B-1347HNH endonuclease [Streptomyces sp. NRRL B-1347]
gi|925387131|gb|LGDA01000250.1|:44655-45964 gi|860602850|ref|WP_048478564.1| gi|860602850|ref|WP_048478564.1| 235 1 235 5.86e-105
LQRGPPALALAALPLLATAPAAHAAGTQAGVPASSGIAGAPGVRAPLPLFEAIDRIPVADEQREGYKWDLYKHWNRGLNATDGCDTRREVILSEAVEAPVVAAGCKMSGGSRLSPYGDVVVTDAAGLDVDHFVPLAEVHDSGSFAWDAVRREAYANDQA
SPLTLIAVTAKSNRSKADKDPAQWIPPASGYHCEYAAQWTAAKLRWSLAADDAERQALLGVAEDCPNTTVVYESAP
MLRGLPALALCTLPLFTAAPAAHSAPTVTAAPAAALTGGPAGLRAPMPLFEAIDRLPVGVEHREGYKRDLYKHWNKGLNAGDGCDTRKEVILAEAVVAPQVAAGCKLTGGSWRSAYDDVVVTDAARLDVDHFVPLAEVYDSEQVPWSAARREAYANDQA
SPDTLIAVSAASNRSKADKDPAEWLPSDGSYHCTYAATWVGTKLRWDLAVDENERQALLGLAEDCPTTSVVYEPAP 998 294 818 Streptomyces roseus HNH endonuclease
[Streptomyces roseus]
gi|925387131|gb|LGDA01000250.1|:44655-45964 gi|664443183|ref|WP_030965733.1| gi|664443183|ref|WP_030965733.1| 233 1 233 2.10e-102
LQRGPPALALAALPLLATAPAAHAAGTQAGVPASSGIAGAP-
GVRAPLPLFEAIDRIPVADEQREGYKWDLYKHWNRGLNATDGCDTRREVILSEAVEAPVVAAGCKMSGGSRLSPYGDVVVTDAAGLDVDHFVPLAEVHDSGSFAWDAVRREAYANDQASPLTLIAVTAKSNRSKADKDPAQWIPPASGYHCEYAAQWTAAKLRWSLAA
DDAERQALLGVAEDCPNTTVVYESAP MRHGLPALTLCALPLLASTPTAHSAPAD---
PVRGPAAGGPAGVRAPLPLYEAIDQLPVAAEHREGYKRDLYKHWNKGLNSGDGCDTRKEVMLAEAVTAPQVAAGCKLSGGSWRSAYDNIVVSDAARLDVDHFVPLAEVYDSEQTPWTAARREAYANDQASPDTLIAVSAASNRSKSDKDPADWLPSDDSYHCTYAATW
VGTKLRWDLAVDENERQALLGLAEDCPTTTVVHETAP 998 294 801 Streptomyces sp. NRRL S-378 HNH endonuclease [Streptomyces sp. NRRL S-378]
gi|925387131|gb|LGDA01000250.1|:44655-45964 gi|926355608|ref|WP_053686546.1| gi|926355608|ref|WP_053686546.1| 232 46 232 1.85e-101
LFEAIDRIPVADEQREGYKWDLYKHWNRGLNATDGCDTRREVILSEAVEAPVVAAGCKMSGGSRLSPYGDVVVTDAAGLDVDHFVPLAEVHDSGSFAWDAVRREAYANDQASPLTLIAVTAKSNRSKADKDPAQWIPPASGYHCEYAAQWTAAKLRWSL
AADDAERQALLGVAEDCPNTTVVYESAP
LFEAIDRIPVADEQRAGYQRNLYKHWNRGLNATDGCDTRREVILAEAVIAPEVAAGCKLTGGSWRSPYDGLLITDAARLDVDHFVPLAEVHDSGGYAWDAARREAYANDQNSPDTLIAVSAASNRSKADKDPAQWLPPYDSYHCTYAATWVGTKLRWSL
AADDAERQALLGLAEDCPTTTVVYEIAP854 294 794 Streptomyces MULTISPECIES: HNH endonuclease [Streptomyces]
gi|925387131|gb|LGDA01000250.1|:44655-45964 gi|664052845|ref|WP_030592179.1| gi|664052845|ref|WP_030592179.1| 241 45 241 6.65e-98
GAPGVRAPLPLFEAIDRIPVADEQREGYKWDLYKHWNRGLNATDGCDTRREVILSEAVEAPVVAAGCKMSGGSRLSPYGDVVVTDAAGLDVDHFVPLAEVHDSGSFAWDAVRREAYANDQASPLTLIAVTAKSNRSKADKDPAQWIPPASGYHCEYAAQ
WTAAKLRWSLAADDAERQALLGVAEDCPNTTVVYESAP
GAGGPLATLPLLDAVERIPVAVEQREGYKRELYKHWNKGLDAGDGCDTRREVIISEAIKAPVVTAGCKLTGGSWLSPYDGVTVSDAGKLDVDHMVPLAEVHDSGGYAWDAGRREAYANDQGSPSTLIAVTAKSNRSKSDKDPAQWLPPAADYRCTYAAE
WTGTKLRWELAADEAERQALLALAGECPTTTVVYGTAP 884 294 772 Streptomyces globisporus HNH endonuclease [Streptomyces globisporus]
gi|925387131|gb|LGDA01000250.1|:44655-45964 gi|1154960293|ref|WP_078865614.1| gi|1154960293|ref|WP_078865614.1| 238 20 238 2.19e-97
LPLLATAPAAHAAGTQAGVPASSGIAGAPGVRAPLPLFEAIDRIPVADEQREGYKWDLYKHWNRGLNATDGCDTRREVILSEAVEAPVVAAGCKMSGGSRLSPYGDVVVTDAAGLDVDHFVPLAEVHDSGSFAWDAVRREAYANDQASPLTLIAVTAKS
NRSKADKDPAQWIPPASGYHCEYAAQWTAAKLRWSLAADDAERQALLGVAEDCPNTTVVYESAP VPAAADTPAAKETTLTQLVP----
VQGAAAVRAPLALFEAIDRLPVAEEHRAGYKRDLYKHWNRGLNATDGCDTRREVILAEAVGAPQVAAGCKLTGGTWRSAYDNLVVTDAARLDVDHFVPLAEVYDSEQTPWSAARREAYANDQNSPDTLIAVSAASNRSKSDKDPAEWLPSDGSYHCTYAATWVGTKLR
WDLAVDENERQALLGLAEDCPTTTVVYEPAP 962 294 768 Streptomyces roseus HNH endonuclease [Streptomyces roseus]
gi|925387131|gb|LGDA01000250.1|:44655-45964 gi|494479287|ref|WP_007268764.1| gi|494479287|ref|WP_007268764.1| 233 41 233 1.87e-96
VRAPLPLFEAIDRIPVADEQREGYKWDLYKHWNRGLNATDGCDTRREVILSEAVEAPVVAAGCKMSGGSRLSPYGDVVVTDAAGLDVDHFVPLAEVHDSGSFAWDAVRREAYANDQASPLTLIAVTAKSNRSKADKDPAQWIPPASGYHCEYAAQWTAA
KLRWSLAADDAERQALLGVAEDCPNTTVVYESAP
LRTPVPLYEAIDRLPVAPEHRDGYRRDLYKHWNRGLNASDGCDTRKEVILSEAVEAPQVAAGCKLTGGSWRSAYDNLVVADAARLDVDHFVPLAEVYDSEQTVWSAARREAYANDQGSPDTLIAVSAASNRSKADKDPSEWLPSDASYHCTYAATWVGT
KLRWNLAVDDAERQALLGLAEDCPTTTVVYETAP 872 294 761 Streptomyces sp. C HNH endonuclease [Streptomyces sp. C]

Spacer 2:

>JQOQ01000126.1_6927_10067_28_spacer_9569_35
AGTCTATTTTGCACTGCTAACGGTTTCAGCATCTC

Hit coordinates:
ContigID: AOUX01000192
Start: 3802
Stop: 3836

Intergenic sequence:
>gi|456831403|gb|AOUX01000192.1|:3205-4402 Leptospira interrogans serovar Naam str. Naam ctg7180000007680, whole genome shotgun sequence
TTAGCTCCGCATCTGTCCAACTTTTTAGGGCGTTTTTTTTAATAGTTTTTTTGGTCTGAATTACGTTGGT
CGCGCGATCAGGATAATATTTTTCAACCAAATCAACCGTATACGGTGGTAAGACTTTTGACTTCTCGTCC
GTAATTTCAATTTCTTGCGAATCGATCCTTTTAATGGTAGCAAAAGGCAATTCAACAGTTTTTTTATTTA
CAAATATTAAATTTTTATTATTCAGAATATATTTACGTATTTGTTTGTCCAAGCTTTCTCTTTCAAAAAC
TAATTCTCGTTCAGTATTTGTAAATTTAGAACGAACAGTTTCAACTTCCGCGTTTTTGTCCCCGTCGATG
CGTTTTAGATCGGATTCGATCTCGATTAATCTTTGAAGTCGTTTATTAAGATCCTCTAATGTTTTAGGCG
GTCGGGTTGAACTCATCGACAGTCTCCTTATTGATTTTTAATCTACTTGCTCGTTCGTTGTCGACTAATA
GTAATTTACGTTTCCCTTTTGGGACCTTTGTTCCCTTTTTTTCGATCGCCTCAATAACAAGGAGCGATTT
AGGTCCTGGGATTAATCTGATTGTTTTATTGTAATCGGAGATGCTGAAACCGTTAGCAGTGCAAAATAGA
CTCATACGGCTATATAATTTGCCTATATGCGCCTTCAGTTTCTTCTCGGTGGCCTCCGAAAGTAATCTTT
CTGACTTTTGGGTTTGTTTGCTCATTTGTACATCCTATAAATTATAATTACTACTATGGTTAGGCATAAT
AAACGTATTATAAAAAGCTCAAGCATCTAGATCCCGAGAAATATAAAACACAGTTACCCAAATAGCAGCA
GCTAAAAAGATCAAGAATGCTTGTAAAAAGAGATACGCTGTCATAACTTTACAACCCCACCAAGTGGATT
AACGGAACCCTGTGATGATGAAGGGCTATAGCGATCAATGAGGCGACGCAGTCTTAAGGCTTGTTTTTTT
GCGTCGGACCGATCCATGTGGAGCTCGTCACCATCACCATCTTTCTGAAAAAATAGATCTGAATGCTCTA
TAGTTTGTTTAAGCACGTCTGAAAAGATCATATATCGGCCTCCACCCAAACTTGTTTATTGTTCCAAATC
TTAACTGCGTCCGCAATTGATTCGCGCAATATTGTGGTTACGCGACAAGAGTGACATCTAACGGCATAGG
CCTTGTTC

Top 10 BLASTX hits: into hypothetical protein

```
# Fields: query id, subject id, subject ids, subject length, s. start, s. end, evalue, query seq, subject seq, q. start, q. end, score, subject sci names, subject title
# 10 hits found
gi|456831403|gb|AOUX01000192.1|:3205-4402      gi|446021850|ref|WP_000099705.1|      gi|446021850|ref|WP_000099705.1|      176      1       145      8.09e-94
         MSSTRPPKTLEDLNKRLQRLIEIESDLKRIDGDKNAEVETVRSKFTNTERELVFERESLDKQIRKYILNNKNLIFVNKKTVELPFATIKRIDSQEIEITDEKSKVLPPYTVDLVEKYYPDRATNVIQTKKTIKKNALKSWTDAEL
         MSSTRPPKTLEDLNKRLQRLIEIESDLKRIDGDKNAEVETVRSKFTNTERELVFERESLDKQIRKYILNNKNLIFVNKKTVELPFATIKRIDSQEIEITDEKSKVLPPYTVDLVEKYYPDRATNVIQTKKTIKKNALKSWTDAEL     436
         2       735      Leptospira interrogans      hypothetical protein [Leptospira interrogans]
gi|456831403|gb|AOUX01000192.1|:3205-4402      gi|488052634|ref|WP_002124031.1|      gi|488052634|ref|WP_002124031.1|      176      1       145      3.59e-93
         MSSTRPPKTLEDLNKRLQRLIEIESDLKRIDGDKNAEVETVRSKFTNTERELVFERESLDKQIRKYILNNKNLIFVNKKTVELPFATIKRIDSQEIEITDEKSKVLPPYTVDLVEKYYPDRATNVIQTKKTIKKNALKSWTDAEL
```

```
        MSSTRPPKTLEDLNKQLQRLIEIESDLKRIDGDKNAEVETVRSKFTNTERELVFERESLDKQIRKYILNNKNLIFVNKKTVELPFATIKRIDSQEIEITDEKSKVLPPYTVDLVEKYYPDRATNVIQTKKTIKKNALKSWTDAEL    436
        2       730     Leptospira interrogans      Gam-like protein [Leptospira interrogans]
gi|456831403|gb|AOUX01000192.1|:3205-4402     gi|516461609|ref|WP_017850447.1|     gi|516461609|ref|WP_017850447.1|     176     1       145     8.23e-87
        MSSTRPPKTLEDLNKRLQRLIEIESDLKRIDGDKNAEVETVRSKFTNTERELVFERESLDKQIRKYILNNKNLIFVNKKTVELPFATIKRIDSQEIEITDEKSKVLPPYTVDLVEKYYPDRATNVIQTKKTIKKNALKSWTDAEL
        MSLTQPPKTLEDLNKRLQRLIKIDSDLKRIDGDKNAEVETVRSKFTDTERDLVFERESLDKQVREYILNNKNLIFVNKKTVELPFATIKRIDSQEIEITDEKSKVLPPYTVDLVEKYYPDRATNVIQTKKTIKKTALKSWTGAEL    436
        2       688     Leptospira interrogans      hypothetical protein [Leptospira interrogans]
gi|456831403|gb|AOUX01000192.1|:3205-4402     gi|490915838|ref|WP_004777741.1|     gi|490915838|ref|WP_004777741.1|     176     1       145     1.75e-79
        MSSTRPPKTLEDLNKRLQRLIEIESDLKRIDGDKNAEVETVRSKFTNTERELVFERESLDKQIRKYILNNKNLIFVNKKTVELPFATIKRIDSQEIEITDEKSKVLPPYTVDLVEKYYPDRATNVIQTKKTIKKNALKSWTDAEL
        MSSTQPPKTLEDLNKRLQRLIEIESDLKRIDGDKNTEVESVRSRFIDTERELVFEKENLDKQVREYILNNKDIIFVHKKTVELPFATIKRIDSQEIEITDEKSKTLPPYTVDLVEKYYSDRAPNVIQIKKTIKKAALKNWTNAEL    436
        2       640     Leptospira kirschneri       Gam-like protein [Leptospira kirschneri]
gi|456831403|gb|AOUX01000192.1|:3205-4402     gi|487897388|ref|WP_001970854.1|     gi|487897388|ref|WP_001970854.1|     178     1       147     1.00e-69
        LSMSSTRPPKTLEDLNKRLQRLIEIESDLKRIDGDKNAEVETVRSKFTNTERELVFERESLDKQIRKYILNNKNLIFVNKKTVELPFATIKRIDSQEIEITDEKSKVLPPYTVDLVEKYYPDRATNVIQTKKTIKKNALKSWTDAEL
        MSTTQNQPPKTLEDLNSKMQRLVEIESDLKRIEGEKNSEVESVRSRFVDTERDLVFEKEELDKQVRDFVMQNKDTLFAHRKTVELPFATIKKIDSQEIEITDEKSKELPPYSVDLIEKFYPERANNAIQIKKTVKKTALKSWTDAEL   442
        2       575     Leptospira interrogans      Gam-like protein [Leptospira interrogans]
gi|456831403|gb|AOUX01000192.1|:3205-4402     gi|1001638526|ref|WP_061216057.1|    gi|1001638526|ref|WP_061216057.1|    176     1       145     4.86e-69
        MSSTRPPKTLEDLNKRLQRLIEIESDLKRIDGDKNAEVETVRSKFTNTERELVFERESLDKQIRKYILNNKNLIFVNKKTVELPFATIKRIDSQEIEITDEKSKVLPPYTVDLVEKYYPDRATNVIQTKKTIKKNALKSWTDAEL
        MNAAQTPKTLEDLNARLQRLVEIEANLKRIGGDKNTEVESVRSRFIDTERDLVFEKEQLDRQVREFVLQNKNTLFVHRKTIELPFATIRKIDSQEIEITDEKSKTLPPYSVDLIEKFYPERASDAIQIKKSIKKAALKSWTDAEL    436
        2       571     Leptospira santarosai        Gam-like protein [Leptospira santarosai]
gi|456831403|gb|AOUX01000192.1|:3205-4402     gi|490641537|ref|WP_004506532.1|     gi|490641537|ref|WP_004506532.1|     178     1       147     1.55e-67
        LSMSSTRPPKTLEDLNKRLQRLIEIESDLKRIDGDKNAEVETVRSKFTNTERELVFERESLDKQIRKYILNNKNLIFVNKKTVELPFATIKRIDSQEIEITDEKSKVLPPYTVDLVEKYYPDRATNVIQTKKTIKKNALKSWTDAEL
        MNRAPNQPPKTLEDLNTKMQRLIEIKSDLKRIEGEKNTEVESIRSRFVGVERDLVFEKEELDKQVREFVMQNKDTLFVHRKTIELPFATIKKIDSQEIEITDEKSKTLPPYSVDLIEKFYPERANNAIQIKKSVKKTALKSWTDAEL   442
        2       561     Leptospira weilii  Gam-like protein [Leptospira weilii]
gi|456831403|gb|AOUX01000192.1|:3205-4402     gi|515129273|ref|WP_016758123.1|     gi|515129273|ref|WP_016758123.1|     177     8       146     1.70e-65
        PKTLEDLNKRLQRLIEIESDLKRIDGDKNAEVETVRSKFTNTERELVFERESLDKQIRKYILNNKNLIFVNKKTVELPFATIKRIDSQEIEITDEKSKVLPPYTVDLVEKYYPDRATNVIQTKKTIKKNALKSWTDAEL
        PKSLEDLNIRMQRLIEIESDLKRIEGEKNTEVEDTRSHFVEVERDLVFEKERLDKQIRAFVMENKNTLFVHRKTIELPFATIRKIDSQEIEITDEKSKTLPPYSVDLIEKFYPERANNAIQIKKSVKKTALKSWTDAEL 418     2
        547     Leptospira          MULTISPECIES: hypothetical protein [Leptospira]
gi|456831403|gb|AOUX01000192.1|:3205-4402     gi|516465861|ref|WP_017854324.1|     gi|516465861|ref|WP_017854324.1|     103     1       103     6.90e-64
        MSKQTQKSERLLSEATEKKLKAHIGKLYSRMSLFCTANGFSISDYNKTIRLIPGPKSLLVIEAIEKKGTKVPKGKRKLLLVDNERASRLKINKETVDEFNPTA
        MSKQTQKSERLLSEATEKKLKAHIGKLYSRMSLFCTANGFSISDYNKTIRLIPGPKSLLVIEAIEKKGTKVPKGKRKLLLVDNERASRLKINKETVDEFNPTA         725     417     529     Leptospira
interrogans         hypothetical protein [Leptospira interrogans]
gi|456831403|gb|AOUX01000192.1|:3205-4402     gi|516474782|ref|WP_017863226.1|     gi|516474782|ref|WP_017863226.1|     176     1       145     1.48e-61
        MSSTRPPKTLEDLNKRLQRLIEIESDLKRIDGDKNAEVETVRSKFTNTERELVFERESLDKQIRKYILNNKNLIFVNKKTVELPFATIKRIDSQEIEITDEKSKVLPPYTVDLVEKYYPDRATNVIQTKKTIKKNALKSWTDAEL
        MNAAQTPKTLEDLNAHLQRLVEIEANLKRIDGDKNTEVESVRSKFIDTERDLVFEKEQLDRQVREFVLQNKDTLFVHRKTIELPFATIRKIDSQEIEITDEKSKTLPPYSVDLIEKFYPERVSDAIQIKKLIKKTTLKSWTDAEL    436
        2       521     Leptospira santarosai        hypothetical protein [Leptospira santarosai]
```

Spacer 3:

>HG916826.1_842637_846204_26_spacer_844192_32
TACATCAGCAAGGACGGCCCGGACATGATCCC

Hit coordinates:
ContigID: LMMB01000004
Start: 84179
Stop: 84210

Intergenic sequence:
>gi|944790866|gb|LMMB01000004.1|:84017-84790 Pseudomonas sp. Leaf83 contig_12, whole genome shotgun sequence
TGTACAAGCCAGTATTTCCGGGGGTTCTAGCCGGATCGCTTCTCAGATGGGAGCGTTTTGGGAACACTTT
GGGAATGGTGGAACGAAAAAGGCAGCCTGCGAGGGCTGCCTTTTGTCGTTTCTAGGGCCTGGATTCACAG
CTTTAGCGCGTGCTCCAGGAGCGGGATCATGTCCGGGCCGTCCTTGCTGATGTACTTGCCGTAATGACGT
TCGAGCATTGCGATTGTCGTGTGCTTCCATACGGCTGACGCAGGCTGGGCAGAGGCACTTGTAACCGCCG
CTGTTGGTGGTCATGGCTGGGGCTTCCTGTGCAGGGCGAGGATGGCGCCCACTTCGGCGTGGCGCGCGGC
GATGTGCTGGCGGTGGGCGGTGATGATGCTGGCCAGCTCGGCTTCGTCGATTTGCCCGTCTTCCAGGGCC
TTGAGGATCATCGCGTCGACCTTGCCGCGCTTTACCGCCGTGCCCATGGCGCGGGCCAGCAGGTCAATGT
TGTCCGTGTCTGTCAGCTCGGGCAGGGGACGTAAACGCCGTTGTAGAGGCTGGTGAGGTAGTCGGGCAGA
TAGCTGGTGCCGGCGACCTTCTCCAGCTGGCGCACTTGTTCGTCCGTCAGCGGGCGGTTGCCGGGGTTCT
CGTACAGCTTGTTGTCGAACTGCTTGAGGTCCAGCCCCGGGCAGGTGGCGGCGCATTCGCGGCCCCCTGG
GAAGGCGCCCCGGTGAGGGGGAATGCGAGGGCGATATCGGCAAGGCCGTGCTGGTGGGCGATCCGTTGTG
AATC

Top 10 BLASTX hits: into hypothetical protein

```
# Fields: query id, subject id, subject ids, subject length, s. start, s. end, evalue, query seq, subject seq, q. start, q. end, score, subject sci names, subject title
# 20 hits found
gi|944790866|gb|LMMB01000004.1|:84017-84790    gi|980973412|ref|WP_059392489.1|     gi|980973412|ref|WP_059392489.1|     156      83       156      1.47e-63
         LPELTDTDNIDLLARAMGTAVKRGKVDAMILKALEDGQIDEAELASIITAHRQHIAARHAEVGAILALHRKPQP MPEPADTDNIDLLARAMGTAVKRGKVDAMILKALEDGQIDEAELASIITAHRQHIAARHAEVGAILALHRKPQP 515
         294      355      Pseudomonas toyotomiensis   hypothetical protein [Pseudomonas toyotomiensis]
gi|944790866|gb|LMMB01000004.1|:84017-84790    gi|980973412|ref|WP_059392489.1|     gi|980973412|ref|WP_059392489.1|     156      18       85       1.47e-63
         GAFPGGRECAATCPGLDLKQFDNKLYENPGNRPLTDEQVRQLEKVAGTSYLPDYLTSLYNGVYVPCPS       AAFPGGRECAATCLGLDLKQFDNKLYENPGHRPLTDEQVRQLEKVAGTSYLPDYITGLYNGVFVAMPE       709
         506      327      Pseudomonas toyotomiensis   hypothetical protein [Pseudomonas toyotomiensis]
gi|944790866|gb|LMMB01000004.1|:84017-84790    gi|545127478|ref|WP_021488921.1|     gi|545127478|ref|WP_021488921.1|     155      82       155      2.10e-63
         LPELTDTDNIDLLARAMGTAVKRGKVDAMILKALEDGQIDEAELASIITAHRQHIAARHAEVGAILALHRKPQP MPELTDTDNIDLLARAMGTAVKRGKVDAMILRALEDGQIDEAELASIITAHRQHIAARHAEVGAILALHRKPQP 515
         294      367      Pseudomonas mendocina       hypothetical protein [Pseudomonas mendocina]
gi|944790866|gb|LMMB01000004.1|:84017-84790    gi|545127478|ref|WP_021488921.1|     gi|545127478|ref|WP_021488921.1|     155      18       84       2.10e-63
         GAFPGGRECAATCPGLDLKQFDNKLYENPGNRPLTDEQVRQLEKVAGTSYLPDYLTSLYNGVYVPCPS       AAFPGGRECAA-CLGLDLKQFDNKLYENPGHRPLTDEQVRQLEKVAGTTYLPDYLTDLYNGVYVAMPE       709
         506      313      Pseudomonas mendocina       hypothetical protein [Pseudomonas mendocina]
gi|944790866|gb|LMMB01000004.1|:84017-84790    gi|917193888|ref|WP_051800600.1|     gi|917193888|ref|WP_051800600.1|     156      83       156      2.14e-62
         LPELTDTDNIDLLARAMGTAVKRGKVDAMILKALEDGQIDEAELASIITAHRQHIAARHAEVGAILALHRKPQP MPEPADTDNIDLLARAMGTAVKRGKVDAMILKALEDGQIDEAELASIITAHRQHIAARHAEVGAILALHRKPQP 515
         294      355      Pseudomonas oleovorans      hypothetical protein [Pseudomonas oleovorans]
gi|944790866|gb|LMMB01000004.1|:84017-84790    gi|917193888|ref|WP_051800600.1|     gi|917193888|ref|WP_051800600.1|     156      18       85       2.14e-62
         GAFPGGRECAATCPGLDLKQFDNKLYENPGNRPLTDEQVRQLEKVAGTSYLPDYLTSLYNGVYVPCPS       AAFPGGRECAAACLGLDLKQLDNKLYENPGHRPLTDEQVRQLEKVAGTSYLPDYITGLYNGVFVAMPE       709
         506      317      Pseudomonas oleovorans      hypothetical protein [Pseudomonas oleovorans]
```

gi|944790866|gb|LMMB01000004.1|:84017-84790 gi|835622920|ref|WP_047589232.1| gi|835622920|ref|WP_047589232.1| 156 83 156 8.82e-60
LPELTDTDNIDLLARAMGTAVKRGKVDAMILKALEDGQIDEAELASIITAHRQHIAARHAEVGAILALHRKPQP MPELADTDNIDLLARAMGTTIKRGTVDAEILKALEDGEISEAELASIIAAHRQHIAARHAEVGAILALHRKPQP 515
294 331 Pseudomonas mendocina hypothetical protein [Pseudomonas mendocina]
gi|944790866|gb|LMMB01000004.1|:84017-84790 gi|835622920|ref|WP_047589232.1| gi|835622920|ref|WP_047589232.1| 156 18 84 8.82e-60
GAFPGGRECAATCPGLDLKQFDNKLYENPGNRPLTDEQVRQLEKVAGTSYLPDYLTSLYNGVYVPCP AAFPGGRECAATCLGLDLKQFDNKLYENPGHRPLTDEQVLQLEKVAGTSYLPDYISGLYNGVFVAMP 709
509 318 Pseudomonas mendocina hypothetical protein [Pseudomonas mendocina]
gi|944790866|gb|LMMB01000004.1|:84017-84790 gi|503480628|ref|WP_013715289.1| gi|503480628|ref|WP_013715289.1| 159 18 84 3.53e-53
GAFPGGRECAATCPGLDLKQFDNKLYENPGNRPLTDEQVRQLEKVAGTSYLPDYLTSLYNGVYVPCP AAFPGGRECAATWLGLDLKQFDNKLYENPGHRPLTDEQVLQLEKVAGTSYLPDYISGLYNGVYVAMP 709
509 313 Pseudomonas mendocina hypothetical protein [Pseudomonas mendocina]
gi|944790866|gb|LMMB01000004.1|:84017-84790 gi|503480628|ref|WP_013715289.1| gi|503480628|ref|WP_013715289.1| 159 83 159 3.53e-53
LPELTDTDNIDLLARAMGTAVKRGKVDAMILKALEDGQIDEAELASIITAHRQHIAARHAEVGAILALHRKPQP*PP
MPELAELDNIDLLERAMTTTIKRGTVDAMILTALKDGEINEAELASIICAHRQHMAARHAEVSSILALHSKRQEPKP 515 285 279 Pseudomonas mendocina hypothetical
protein [Pseudomonas mendocina]

Spacer 4:

>LFQC01000004.1_87106_87459_1_spacer_87136_35
GTAACAGATGATATTAAATTCAAAAGAATTAACTT

Hit coordinates:
ContigID: AOSX01000029
Start: 620311
Stop: 620345

Intergenic sequence:
>gi|727535586|gb|AOSX01000029.2|:619946-620968 Clostridium botulinum Af84 Contig_29, whole genome shotgun sequence
AAATTCGGATAGAAACATAGATATTTCTCATATAGATTTTTTTGATTAATTATATGCACTGATGCTATGA
ATAAAAACAAAAAAATTCTAAAAGTTTTTCCAACCGCATTTAATTTTAAAACCTATCCTAAGCATTGCAA
TATAAGGCTTAAGATAGGTTTTGTTTATGTTTTTCTGAAAAATCACAAATGGCTGGAAAAATTTTTTTAA
TCATTGTAAAATCAAATGCTATATGCTATTCTTAAAATTAAGGAATGGCTAATTTACTATGGATGAACAT
TAACATAGGATGTATTTAAATTTGGAAAGCTTACTAGATTTCTTACCTATATCGGGTTGAACAATAACAT
GAGATGCATTTAAATAAGTTAATTCTTTTGAATTTAATATCATCTGTTACACTGAACAATAACATAAGAT
GTATTATTAAAAATGGTAGTTGTAGCTTATGTTATAGCTATTATTTTTCAATATTTAGAGGAATATTTTT
ACATATGTAGAATATTTGTATATAGTCCCTTTAATTCATTATCAAATCTTTCTTGAATAAAAAGGAGCCT
GTGTATGCAGTGGGGGCTCCTTTTTTATATTATTTGAAGGAATTTTTTAACATTTATAGAATATTAAATA
TAACGGCTTCCCAATAGGTTGATTATAGACCTCCTTGCTTAAGCAAAAAGAACCCCAATAAAAAGGGGTT
CTTTTTGTATGGATTTATATATTGGTGCTATGTATTGGTCTATTTTTACTATATCCAGATGTAGAAAAGT
TAATCAGTAAGAAATATATAAGATGTAAATTTTTAAATGTATTTCTATAATGAGAGTACCTTATTTTTTA
TTTTTTTGATAAAAATTTTAAGAAAATTAGCAATTTATTAATAAAAATTTAAAATAATTATTTAATTTCA
TTTTAAAGTTGTATCAGAATGTATAAAAAAGTAAAAAGATGTAGGATAAACAAAGAAGGAATTGAATTAA
TTGTAAAGAATATAAATATTATGGTAAAAATATAAAATAAATA

Top 10 BLASTX hits: (1 hit found) into hypothetical protein
Fields: query id, subject id, subject ids, subject length, s. start, s. end, evalue, query seq, subject seq, q. start, q. end, score, subject sci
names, subject title
1 hits found
gi|727535586|gb|AOSX01000029.2|:619946-620968 gi|1119667520|ref|WP_072587154.1| gi|1119667520|ref|WP_072587154.1| 80 49
79 2.91e-07VTLNNNIRCIIKNGSCSLCYSYYFSIFRGIF LKLNNNIRCIIKNGSCSLCYSYYFSLFRRIF 396 488 136 Clostridium botulinum
hypothetical protein [Clostridium botulinum]

Supplementary text 3
Examples of mispredicted CRISPR arrays

Predicted CRISPR-Array: 1

ContigID: AKOR01000006.1 (Helicobacter_pylori_Hp_A_6_GCA_000273985.1)
Start: 34315
Stop: 34512
https://www.ncbi.nlm.nih.gov/projects/sviewer/?id=AKOR01000006.1&v=34315..34513
Predicted in: putative vacuolating cytotoxin family protein

CRISPR-Finder prediction:
CRISPR id : tmp_1_Crispr_1

- CRISPR start position : 101 ---------- CRISPR end position : 298 ---------- CRISPR length : 197
- DR consensus : AACAGCAATTTCAATCAAGGGACTTAT
- DR length : 27 Number of spacers : 3

101	**CAAACCAGCTTCAATCAAGGGACTTAT** AATTTTAGTAACAGCGCCACTTTAAGCTTCAAT	160
161	**AACAGCAATTTCAATCAAGGGACTTAT** CACTTTAACAGCGCCCAATCCACTTTTGAA	217
218	**AACAGCAATTTCAATCAAGGCACTTAT** AATTTTAATGACAATACTAGCTTTAAT	271
272	**AACGACACCTTCAATCAAGGCGCTTAT**	298

Top 10 BLASTX results in RefSeq:
Fields: query id, subject id, subject ids, subject length, s. start, s. end, evalue, query seq, subject seq, q. start, q. end, score, subject sci names, subject title
10 hits found

```
gi|393068133|gb|AKOR01000006.1|:34315-34513      gi|446797347|ref|WP_000874603.1|
        gi|446797347|ref|WP_000874603.1| 2662     258      323      5.45e-32
        QTSFNQGTYNFSNSATLSFNNSNFNQGTYHFNSAQSTFENSNFNQGTYNFNDNTSFNNDTFNQGAY
        QTSFNQGTYNFSNSATLSFNNSNFNQGTYHFNSAQSTFENSNFNQGTYNFNDNTSFNNDTFNQGAY        1       198      313
        Helicobacter pylori      toxin outer membrane protein [Helicobacter pylori]
gi|393068133|gb|AKOR01000006.1|:34315-34513      gi|1039496436|ref|WP_064832845.1|
        gi|1039496436|ref|WP_064832845.1|         2528     258      323      2.71e-31
        QTSFNQGTYNFSNSATLSFNNSNFNQGTYHFNSAQSTFENSNFNQGTYNFNDNTSFNNDTFNQGAY
        QTSFNQGTYNFSNSATLSFNNSNFNQGTYHFNSAQSTFENSNFNQGTYNFNDNTSFNNDTFNQGTY        1       198      308
        Helicobacter pylori      toxin [Helicobacter pylori]
gi|393068133|gb|AKOR01000006.1|:34315-34513      gi|1039479039|ref|WP_064816966.1|
        gi|1039479039|ref|WP_064816966.1|         2530     259      324      2.71e-31
        QTSFNQGTYNFSNSATLSFNNSNFNQGTYHFNSAQSTFENSNFNQGTYNFNDNTSFNNDTFNQGAY
        QTSFNQGTYNFSNSATLSFNNSNFNQGTYHFNSAQSTFENSNFNQGTYNFNDNTSFNNDTFNQGTY        1       198      308
        Helicobacter pylori      toxin [Helicobacter pylori]
gi|393068133|gb|AKOR01000006.1|:34315-34513      gi|446797335|ref|WP_000874591.1|
        gi|446797335|ref|WP_000874591.1| 2399     258      323      2.93e-31
        QTSFNQGTYNFSNSATLSFNNSNFNQGTYHFNSAQSTFENSNFNQGTYNFNDNTSFNNDTFNQGAY
        QTSFNQGTYNFSNSATLSFNNSNFNQGTYHFNSAQSTFENSNFNQGTYNFNDNTSFNNDTFNQGTY        1       198      308
        Helicobacter pylori      toxin outer membrane protein [Helicobacter pylori]
gi|393068133|gb|AKOR01000006.1|:34315-34513      gi|446797351|ref|WP_000874607.1|
        gi|446797351|ref|WP_000874607.1| 2399     258      323      2.93e-31
        QTSFNQGTYNFSNSATLSFNNSNFNQGTYHFNSAQSTFENSNFNQGTYNFNDNTSFNNDTFNQGAY
        QTSFNQGTYNFSNSATLSFNNSNFNQGTYHFNSAQSTFENSNFNQGTYNFNDNTSFNNDTFNQGTY        1       198      308
        Helicobacter pylori      toxin outer membrane protein [Helicobacter pylori]
gi|393068133|gb|AKOR01000006.1|:34315-34513      gi|487847449|ref|WP_001920915.1|
        gi|487847449|ref|WP_001920915.1| 2525     258      323      3.19e-31
        QTSFNQGTYNFSNSATLSFNNSNFNQGTYHFNSAQSTFENSNFNQGTYNFNDNTSFNNDTFNQGAY
        QTSFNQGTYNFSNSATLSFNNSNFNQGTYHFNSAQSTFENSNFNQGTYNFNDNTSFNNDTFNQGTY        1       198      307
        Helicobacter pylori      vacuolating cytotoxin family protein [Helicobacter pylori]
```

gi|393068133|gb|AKOR01000006.1|:34315-34513 gi|446797338|ref|WP_000874594.1|
gi|446797338|ref|WP_000874594.1| 2399 258 323 3.26e-31
QTSFNQGTYNFSNSATLSFNNSNFNQGTYHFNSAQSTFENSNFNQGTYNFNDNTSFNNDTFNQGAY
QTSFNQGTYNFSNSATLSFNNSNFNQGTYHFNSAQSTFENSNFNQGTYNFNDNTSFNNDTFNQGTY 1 198 307
Helicobacter pylori toxin-like outer membrane protein [Helicobacter pylori]
gi|393068133|gb|AKOR01000006.1|:34315-34513 gi|446797381|ref|WP_000874637.1|
gi|446797381|ref|WP_000874637.1| 2529 258 323 1.63e-30
QTSFNQGTYNFSNSATLSFNNSNFNQGTYHFNSAQSTFENSNFNQGTYNFNDNTSFNNDTFNQGAY
QTSFNQGTYNFSNSATLSFNNSNFNQGTYHFNSTQSTFENSNFNQGTYNFNDNTSFNNDTFNQGTY 1 198 302
Helicobacter pylori toxin outer membrane protein [Helicobacter pylori]
gi|393068133|gb|AKOR01000006.1|:34315-34513 gi|446797342|ref|WP_000874598.1|
gi|446797342|ref|WP_000874598.1| 2399 258 323 2.63e-30
QTSFNQGTYNFSNSATLSFNNSNFNQGTYHFNSAQSTFENSNFNQGTYNFNDNTSFNNDTFNQGAY
QTSFNQGTYDFSNSATLSFNNSNFNQGTYHFNSAQSTFENSNFNQGAYNFNDNTSFNNDTFNQGTY 1 198 301
Helicobacter pylori toxin outer membrane protein [Helicobacter pylori]
gi|393068133|gb|AKOR01000006.1|:34315-34513 gi|1159760846|ref|WP_079307344.1|
gi|1159760846|ref|WP_079307344.1| 2397 259 324 3.32e-30
QTSFNQGTYNFSNSATLSFNNSNFNQGTYHFNSAQSTFENSNFNQGTYNFNDNTSFNNDTFNQGAY
QTSFNQGTYNFSNSATLSFGNSNFNQGTYHFNSAQSTFENSNFNQGTYNFNDNVSFNNDTFNQGTY 1 198 300
Helicobacter pylori toxin [Helicobacter pylori]

Predicted CRISPR-Array: 2

ContigID: KK099646.1 (Staphylococcus_aureus_DAR3163_GCA_000610585.1)
Start: 60277
Stop: 60684
https://www.ncbi.nlm.nih.gov/projects/sviewer/?id=KK099646.1&v=60277..60684
Predicted in: serine-aspartate repeat-containing protein

CRISPR-Finder prediction:

CRISPR id : tmp_1_PossibleCrispr_1

- CRISPR start position : 1 ---------- CRISPR end position : 132 ---------- CRISPR length : 131
- DR consensus : AGCGATTCAGACTCAGATAGCGACTCAGA
- DR length : 29 Number of spacers : 2

1 **AGTGACTCAGATTCGGACAGCGATTCAGA** CTCAGATAGCGACTCAGATTCAGAT 54
55 **AGTGACTCAGACTCAGATAGCGACTCAGA** CTCAGATAGCGACTCAGAC 102
103 **TCAGATAGCGACTCAGACAGCGACTCAGA** 132

CRISPR id : tmp_1_PossibleCrispr_2

- CRISPR start position : 278 ---------- CRISPR end position : 408 ---------- CRISPR length : 130
- DR consensus : AGCGATTCAGACTCAGATAGCGACTCAGA
- DR length : 29 Number of spacers : 2

278 **AGCGATTCAGATTCAGACAGCGACTCAGA** TTCAGATAGCGACTCAGACTCAGAC 331
332 **AGCGATTCAGACTCAGATAGCGACTCAGA** CAGCGATTCAGATTCGGAT 379
380 **AGCGATTCAGATTCAGATGCAGGTAAACA** 408

Top 10 BLASTX results in RefSeq:
Fields: query id, subject id, subject ids, subject length, s. start, s. end, evalue, query seq, subject seq, q. start, q. end, score, subject sci names, subject title
587 hits found
gi|601598167|gb|KK099646.1|:60277-60684 gi|1093956063|ref|WP_070956199.1|
gi|1093956063|ref|WP_070956199.1| 1426 1225 1281 7.54e-09
SDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDS
SDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDS 1 171 150 Staphylococcus aureus hydrolase [Staphylococcus aureus]
gi|601598167|gb|KK099646.1|:60277-60684 gi|1093956063|ref|WP_070956199.1|
gi|1093956063|ref|WP_070956199.1| 1426 1227 1283 7.54e-09
SDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDS
SDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDS 1 171 150 Staphylococcus aureus hydrolase [Staphylococcus aureus]
gi|601598167|gb|KK099646.1|:60277-60684 gi|1093956063|ref|WP_070956199.1|
gi|1093956063|ref|WP_070956199.1| 1426 1229 1285 7.54e-09
SDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDS
SDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDS 1 171 150 Staphylococcus aureus hydrolase [Staphylococcus aureus]
gi|601598167|gb|KK099646.1|:60277-60684 gi|1093956063|ref|WP_070956199.1|
gi|1093956063|ref|WP_070956199.1| 1426 1231 1287 7.54e-09
SDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDS
SDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDS 1 171 150 Staphylococcus aureus hydrolase [Staphylococcus aureus]

gi|601598167|gb|KK099646.1|:60277-60684 gi|1093956063|ref|WP_070956199.1|
gi|1093956063|ref|WP_070956199.1| 1426 1233 1289 7.54e-09
SDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDS
SDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDS 1 171 150 Staphylococcus aureus hydrolase [Staphylococcus aureus]
gi|601598167|gb|KK099646.1|:60277-60684 gi|1093956063|ref|WP_070956199.1|
gi|1093956063|ref|WP_070956199.1| 1426 1235 1291 7.54e-09
SDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDS
SDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDS 1 171 150 Staphylococcus aureus hydrolase [Staphylococcus aureus]
gi|601598167|gb|KK099646.1|:60277-60684 gi|1093956063|ref|WP_070956199.1|
gi|1093956063|ref|WP_070956199.1| 1426 1237 1293 7.54e-09
SDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDS
SDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDS 1 171 150 Staphylococcus aureus hydrolase [Staphylococcus aureus]
gi|601598167|gb|KK099646.1|:60277-60684 gi|1093956063|ref|WP_070956199.1|
gi|1093956063|ref|WP_070956199.1| 1426 1239 1295 7.54e-09
SDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDS
SDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDS 1 171 150 Staphylococcus aureus hydrolase [Staphylococcus aureus]
gi|601598167|gb|KK099646.1|:60277-60684 gi|1093956063|ref|WP_070956199.1|
gi|1093956063|ref|WP_070956199.1| 1426 1241 1297 7.54e-09
SDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDS
SDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDS 1 171 150 Staphylococcus aureus hydrolase [Staphylococcus aureus]
gi|601598167|gb|KK099646.1|:60277-60684 gi|1093956063|ref|WP_070956199.1|
gi|1093956063|ref|WP_070956199.1| 1426 1243 1299 7.54e-09
SDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDS
SDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDS 1 171 150 Staphylococcus aureus hydrolase [Staphylococcus aureus]

Predicted CRISPR-Array: 3

ContigID: ANAE01000140.1 (Nocardiopsis_prasina_DSM_43845_GCA_000341265.1)
Start: 57689
Stop: 58634
https://www.ncbi.nlm.nih.gov/projects/sviewer/?id=ANAE01000140.1&v=57689..58634

Array predicted by PilerCR:
>gi|451810799|gb|ANAE01000140.1|:57689-58634 Nocardiopsis prasina DSM 43845 contig_140, whole genome shotgun sequence
CCGGTCCATCCCCGCGTGCGCGGGGCTCACGCGCTGGCCATGGCCGGTGGGTGCGTGCTGGTCGGTCCAT
CCCCGCGTGCGCGGGGCTCACTCGTGCCTTGGCGTTGGTGGGGACTTCGCCACCGGTCCATCCCCGCGTG
CGCGGGGCTCACCCACATCTTGGCGTTGGTGGGGACTTCGCCACCGGTCCATCCCCGCGTGCGCGGGGCT
CACGTCATCGGACGGTGCTCAGCGGGTGCGCGAGCCGGTCCATCCCCGCGTGCGCGGGGCTCACGTCATC
CACATGGCCACCACCTACACCTACACCGGTCCATCCCCGCGTGCGCGGGGCTCACCCACATCTTGGCGTT
GGTGGGGACTTCGCCACCGGTCCATCCCCGCGTGCGCGGGGCTCACGCACATCTTGGCGTTGGTGGGGAC
TTCGCCACCGGTCCATCCCCGCGTGCGCGGGGCTCACCCACATCTTGGCGTTGGTGGGGACTTCGCCACC
GGTCCATCCCCGCGTGCGCGGGGCTCACCCACATCTTGGCGTTGGTGGGGACTTCGCCACCGGTCCATCC
CCGCGTGCGCGGGGCTCACCCACATCTTGGCGTTGGTGGGGACTTCGCCACCGGTCCATCCCCGCGTGCG
CGGGGCTCACCCACATCTTGGCGTTGGTGGGGACTTCGCCACCGGTCCATCCCCGCGTGCGCGGGGCTCA
CCCACATCTTGGCGTTGGTGGGGACTTCGCCACCGGTCCATCCCCGCGTGCGCGGGGCTCACCCACATCT
TGGCGTTGGTGGGGACTTCGCCACCGGTCCATCCCCGCGTGCGCGGGGCTCACCCACATCTTGGCGTTGG
TGGGGACTTCGCCACCGGTCCATCCCCGCGTGCGCGGGGCTCACCCACATCTTGGCGTTGGTGGGGACTT
CGCCACCGGTCCATCCCCGCGTGCGCGGGGCTCACG

Top 10 BLASTX results in RefSeq:
Fields: query id, subject id, subject ids, subject length, s. start, s. end, evalue, query seq, subject seq, q. start, q. end, score, subject sci names, subject title
6 hits found

```
gi|451810799|gb|ANAE01000140.1|:57689-58634      gi|485835026|ref|WP_001446733.1|
        gi|485835026|ref|WP_001446733.1|;gi|218706248|ref|YP_002413767.1| 172     2       167     1.36e-11
        RCSAGARAGP--SPRARGSRHPHGHHLHLHRSIPACAGLTHILALVGTSPPVHPRVRGAHAHLGVGGDFATGPSPRARGSP-----
TSWRWWGLRHRSIPACAGLT--HILALVGTSPPVHPRVRGAHPHLGVGGDFATGPSPRARGSPTSWRWWGLRHRSIPACAGLTHILA
        KWSSTATVGPGLSPLARGTRWWRGQTEQSRRFIPAGAGNTLTYNTENHTLTVYPRWRGEHGHSPKQTQEVNGLSPLARGTPGFHLSKCWMW-----
RFIPAGAGNTTPDIFATVVT--PVYPRWRGEHRNQTQRANGVAGLSPLARGTLALFLSGESRHRFIPAGAGNTYFLS     223     714     174
        Escherichia coli;Escherichia coli UMN026 hypothetical protein [Escherichia coli]
gi|451810799|gb|ANAE01000140.1|:57689-58634      gi|485835026|ref|WP_001446733.1|
        gi|485835026|ref|WP_001446733.1|;gi|218706248|ref|YP_002413767.1| 172     12      167     2.32e-10
        GPSPRARGSRAGHGRWVRAGRSIPACAGLTRALALVGTSPPVHPRVRGAHPHLGVGGDFATGPSPRARGSRHRTVLSGCASRSIPACAGLTSSTWPPP
TPTPVHPRVRGAHPHLGVGGDFATGPSPRARGSRTSWRWWGLRHRSIPACAGLTHILA
        GLSPLARGTRWWRGQTEQSRRFIPAGAGNTLTYNTENHTLTVYPRWRGEHGHSPKQTQEVNGLSPLARGTPGFHLSKCWMWRFIPAGAGNTTPDIFAT
VVTPVYPRWRGEHRNQTQRANGVAGLSPLARGTLALFLSGESRHRFIPAGAGNTYFLS       3       470     165     Escherichia
coli;Escherichia coli UMN026     hypothetical protein [Escherichia coli]
gi|451810799|gb|ANAE01000140.1|:57689-58634      gi|485835026|ref|WP_001446733.1|
        gi|485835026|ref|WP_001446733.1|;gi|218706248|ref|YP_002413767.1| 172     12      167     1.15e-09
        GPSPRARGSPTSWRWW----GLRHRSIPACAGLTHILALVGTSPPVHPRVRGAHPHLGVGGDFATGPSPRARGSP-----
TSWRWWGLRHRSIPACAGLT--HILALVGTSPPVHPRVRGAHPHLGVGGDFATGPSPRARGSPTSWRWWGLRHRSIPACAGLTHILA   GLSPLARGT----
RWWRGQTEQSRRFIPAGAGNTLTYNTENHTLTVYPRWRGEHGHSPKQTQEVNGLSPLARGTPGFHLSKCWMW-----RFIPAGAGNTTPDIFATVVT--
PVYPRWRGEHRNQTQRANGVAGLSPLARGTLALFLSGESRHRFIPAGAGNTYFLS  430     897     160     Escherichia
coli;Escherichia coli UMN026     hypothetical protein [Escherichia coli]
gi|451810799|gb|ANAE01000140.1|:57689-58634      gi|485835026|ref|WP_001446733.1|
        gi|485835026|ref|WP_001446733.1|;gi|218706248|ref|YP_002413767.1| 172     12      167     1.82e-08
        GPSPRARGSLVPWRWW----GLRHRSIPACAGLTHILALVGTSPPVHPRVRGAHVIGRCSAGARAGPSPRARGSRHPHGHHLHLHRSIPACAGLT--
HILALVGTSPPVHPRVRGAHAHLGVGGDFATGPSPRARGSPTSWRWWGLRHRSIPACAGLTHILA GLSPLARGT----
RWWRGQTEQSRRFIPAGAGNTLTYNTENHTLTVYPRWRGEHGHSPKQTQEVNGLSPLARGTPGFHLSKCWMWRFIPAGAGNTTPDIFATVVT--
PVYPRWRGEHRNQTQRANGVAGLSPLARGTLALFLSGESRHRFIPAGAGNTYFLS  64      531     151     Escherichia
coli;Escherichia coli UMN026     hypothetical protein [Escherichia coli]
gi|451810799|gb|ANAE01000140.1|:57689-58634      gi|485835026|ref|WP_001446733.1|
        gi|485835026|ref|WP_001446733.1|;gi|218706248|ref|YP_002413767.1| 172     12      163     3.34e-08
        GPSPRARGSPTSWRWW----GLRHRSIPACAGLTHILALVGTSPPVHPRVRGAHPHLGVGGDFATGPSPRARGSP-----
TSWRWWGLRHRSIPACAGLT--HILALVGTSPPVHPRVRGAHPHLGVGGDFATGPSPRARGSPTSWRWWGLRHRSIPACAGLT       GLSPLARGT----
RWWRGQTEQSRRFIPAGAGNTLTYNTENHTLTVYPRWRGEHGHSPKQTQEVNGLSPLARGTPGFHLSKCWMW-----RFIPAGAGNTTPDIFATVVT--
PVYPRWRGEHRNQTQRANGVAGLSPLARGTLALFLSGESRHRFIPAGAGNT      491     946     149     Escherichia
coli;Escherichia coli UMN026     hypothetical protein [Escherichia coli]
gi|451810799|gb|ANAE01000140.1|:57689-58634      gi|485668974|ref|WP_001310069.1|
        gi|485668974|ref|WP_001310069.1|;gi|218706249|ref|YP_002413768.1| 126     12      122     6.39e-06
        GPSPRARGSRAGHGRWVRAGRSIPACAGLTRALALVGTSPPVHPRVRGAHPHLGVGGDFATGPSPRARGSRHRTVLSGCASRSIPACAGLTSSTWPPP
TPTPVHPRVRGAH
        GLSPLARGTRIFYRRMSRGNRFIPAGAGNTSAAPANVSAVTVYPRWRGEHPRPSIISCASFGLSPLARGTHQEADESARHARFIPAGAGNTQFHMSRR
AGSSVYPRWRGEH   3       335     130     Escherichia coli;Escherichia coli UMN026 hypothetical protein
[Escherichia coli]
# BLAST processed 1 queries
```

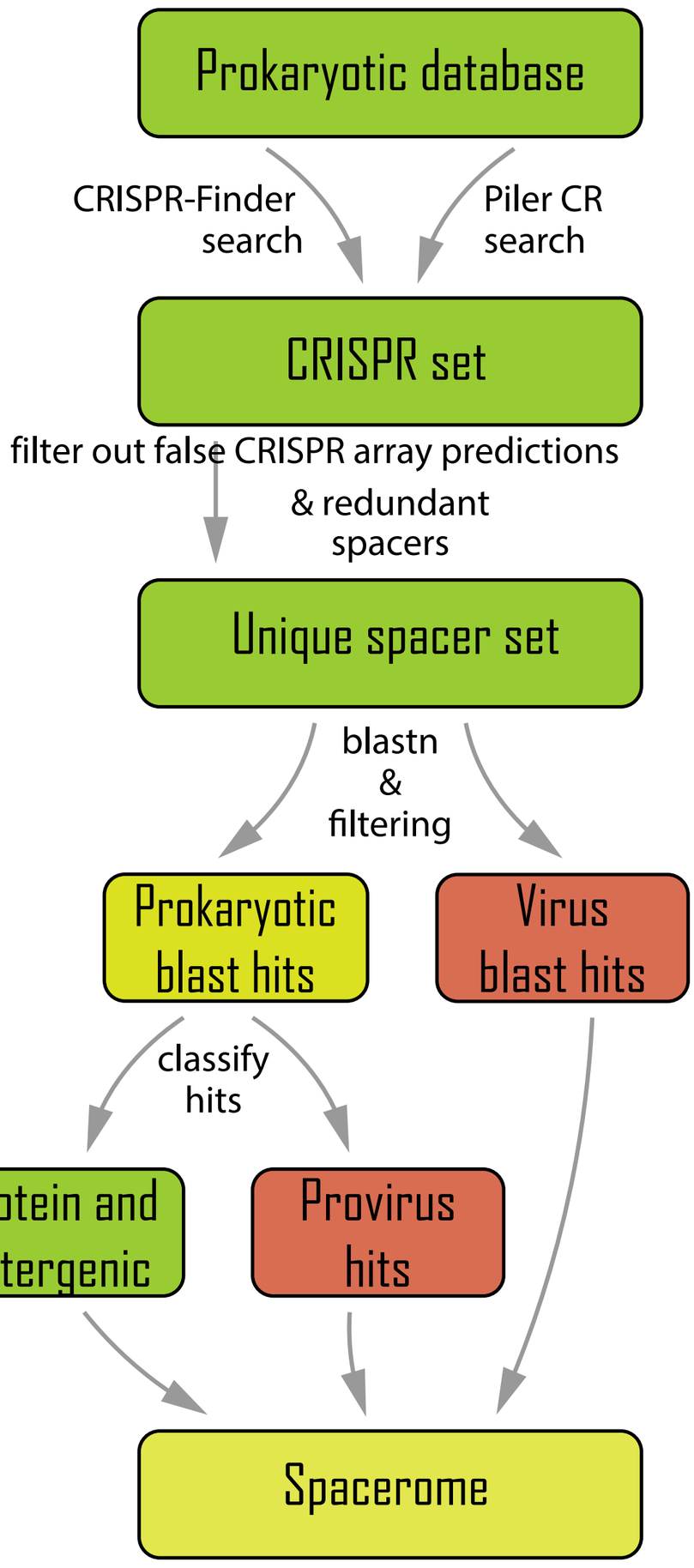
Predicted CRISPR-Array: 4

ContigID: KQ087587.1 (Enterobacter_aerogenes_UCI97_GCA_001030185.1)
Start: 84822
Stop: 86069
https://www.ncbi.nlm.nih.gov/projects/sviewer/?id=KQ087587.1&v=84822..86069

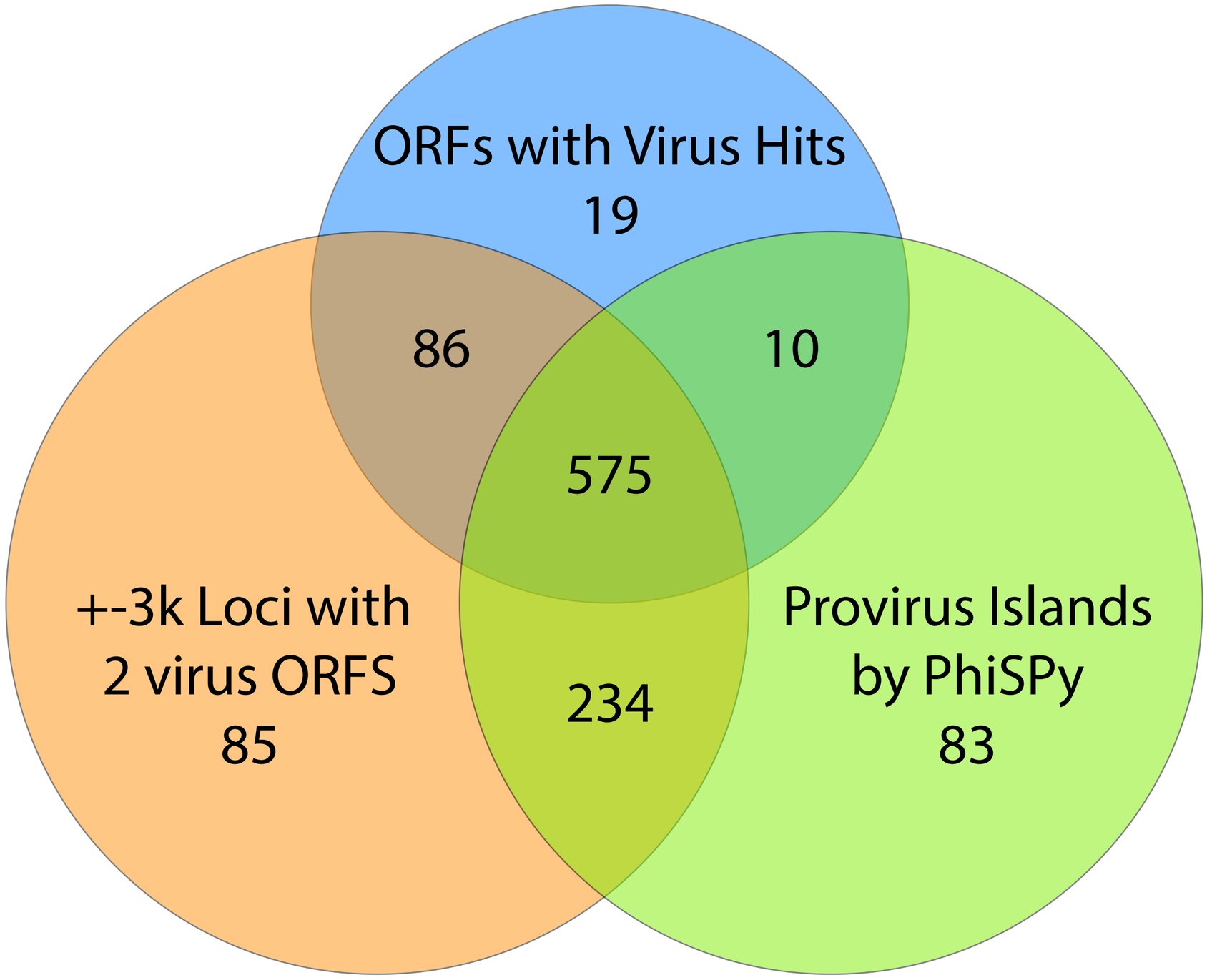
PilerCR predicted Array
>gi|844486473|gb|KQ087587.1|:84822-86069 Enterobacter aerogenes strain UCI97 genomic scaffold aeuqd-supercont1.1, whole genome shotgun sequence
GCGCCTTATCCGGGCTACGGTTCGGTATTTGGTTGGTAGCCCCGGTAAGCGTAAGCGCCACCGGGGGAGG
ATTCCCGGATAGCGGCGCGTAGCGCCTTATCCGGGCTACGGTTCGGTATGCGGTTGGTAGCCCCGGTAAG
CGTAAGCGCCACCGGGGGAGGGTTCCCGGATGGCGGCGCATAGCGCCTTATCCGGGCTACGGTTCGGTAG
TCGGTTGGTAGCCCCGGTAAGCGTAAGCGCCACCGGGGAGGGTTCCCGGATGGCGGCGCATGGCGCCTTA
TCCGGGCTACGGTTCGGTATTCGGTCGGTAGCCCCGGTAAGCGTAAGCGCCACCGGGGAGGATTCCCGGA
TAGCGGCGCATCGCGCCTTATCCGGGCTACGATTCGGTATTCGGTTGGTAGCTCCGGTAAGCGTAAGCGC
CACCGGGGAGGGTTCCCGGATGGCGGCGCNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNTAAAACTGCGC
GCGTGGGTCGCCAGCTTCGGCCTGCTGTTGGTATTGGCTGCCCGGATCCGCAGCGTCTGCGGCGAAGCTT
GTCAGCGCCAGCAGTAAATAGAGGTATTTCATTTTTTTATTATCAATAAGGGAGACCCCCGCAGTATAAA
GGGCGGCGGAAAAGAAGAAAATGGCGGGAATTAGCGGGGTGAATACCCGGATAGCGGCGCATAGCGCCTT
ATCCGGGCTACGGTTCGGTATTTGGTTGGTAGCCCCGGTAAGCGTAAGCGCCACCGGGGGAGGATTCCCG
GATAGCGGCGCGTAGCGCCTTATCCGGGCTACGGTTCGGTATGCGGTTGGTAGCCCCGGTAAGCGTAAGC
GCCACCGGGGGAGGGTTCCCGGATGGCGGCGCATAGCGCCTTATCCGGGCTACGGTTCGGTAGTCGGTTG
GTAGCCCCGGTAAGCGTAAGCGCCACCGGGGAGGGTTCCCGGATGGCGGCGCATGGCGCCTTATCCGGGC
TACGGTTCGGTATTCGGTCGGTAGCCCCGGTAAGCGTAAGCGCCACCGGGGAGGATTCCCGGATAGCGGC
GCATCGCGCCTTATCCGGGCTACGATTCGGTATTCGGTTGGTAGCTCCGGTAAGCGTAAGCGCCACCGGG
GAGGGTTCCCGGATGGCGGCGCATAGCGCCTTATCCGGGCTACGGTTCGGTAGTCGGT

Top 10 BLASTX results in RefSeq:
Fields: query id, subject id, subject ids, subject length, s. start, s. end, evalue, query seq, subject seq, q. start, q. end, score, subject sci names, subject title
99 hits found
gi|844486473|gb|KQ087587.1|:84822-86069 gi|1172898100|ref|WP_080967648.1|
gi|1172898100|ref|WP_080967648.1| 135 11 135 2.47e-77
RLPNRSPDKALCAAIREPSPVALTLTGATNRIPNRSPDKARCAAIRESSPVALTLTGATDRIPNRSPDKAPCAAIREPSPVALTLTGATNRLPNRSPD
KALCAAIREPSPGGAYAYRGYQPHTEP
RLPNRSPDKALCAAIREPSPVALTLTGATNRIPNRSPDKARCAAIRESSPVALTLTGATDRIPNRSPDKAPCAAIREPSPVALTLTGATNRLPNRSPD
KALCAAIREPSPGGAYAYRGYQPHTEP 1246 872 623 Enterobacter aerogenes hypothetical protein [Klebsiella aerogenes]
gi|844486473|gb|KQ087587.1|:84822-86069 gi|1172898100|ref|WP_080967648.1|
gi|1172898100|ref|WP_080967648.1| 135 23 135 2.63e-68
AAIREPSPVALTLTGATNRIPNRSPDKARCAAIRESSPVALTLTGATDRIPNRSPDKAPCAAIREPSPVALTLTGATNRLPNRSPDKALCAAIREPSP
GGAYAYRGYQPHTEP
AAIREPSPVALTLTGATNRIPNRSPDKARCAAIRESSPVALTLTGATDRIPNRSPDKAPCAAIREPSPVALTLTGATNRLPNRSPDKALCAAIREPSP
GGAYAYRGYQPHTEP 447 109 563 Enterobacter aerogenes hypothetical protein [Klebsiella aerogenes]
gi|844486473|gb|KQ087587.1|:84822-86069 gi|1172898100|ref|WP_080967648.1|
gi|1172898100|ref|WP_080967648.1| 135 11 135 1.27e-18
RLPELPTEYRIVARIRRDAPLSGNPPRWRLRLPGLPTEYRTVARIRRHAPPSGNPPRWRLRLPGLPTDYRTVARIRRYAPPSGNPPPVALTLTGATNR
IPNRSPDKALRAAIRESSPGGAYAYRGYQPNTEP
RLPNRSPDKALCAAIREPSPVALTLTGATNRIPNRSPDKARCAAIRESSPVALTLTGATDRIPNRSPDKAPCAAIRE-------
PSPVALTLTGATNRLPNRSPDKALCAAIREPSPGGAYAYRGYQPHTEP 413 18 225 Enterobacter aerogenes
hypothetical protein [Klebsiella aerogenes]
gi|844486473|gb|KQ087587.1|:84822-86069 gi|1172898100|ref|WP_080967648.1|
gi|1172898100|ref|WP_080967648.1| 135 11 135 1.60e-18
RLPELPTEYRIVARIRRDAPLSGNPPRWRLRLPGLPTEYRTVARIRRHAPPSGNPPRWRLRLPGLPTDYRTVARIRRYAPPSGNPPPVALTLTGATNR
IPNRSPDKALRAAIRESSPGGAYAYRGYQPNTEP
RLPNRSPDKALCAAIREPSPVALTLTGATNRIPNRSPDKARCAAIRESSPVALTLTGATDRIPNRSPDKAPCAAIRE-------
PSPVALTLTGATNRLPNRSPDKALCAAIREPSPGGAYAYRGYQPHTEP 1176 781 224 Enterobacter aerogenes
hypothetical protein [Klebsiella aerogenes]
gi|844486473|gb|KQ087587.1|:84822-86069 gi|1172898100|ref|WP_080967648.1|
gi|1172898100|ref|WP_080967648.1| 135 27 61 1.06e-13

```
        NPPPVALTLTGATNQIPNRSPDKALCAAIRVFTPL      EPSPVALTLTGATNRIPNRSPDKARCAAIRESSPV      836     732
        130     Enterobacter aerogenes   hypothetical protein [Klebsiella aerogenes]
gi|844486473|gb|KQ087587.1|:84822-86069  gi|1172898100|ref|WP_080967648.1|
        gi|1172898100|ref|WP_080967648.1|        135     1       30      1.06e-13
        VALTLTGATNRIPNRSPDKALRAAIRESSP   MALTLTGATNRLPNRSPDKALCAAIREPSP   915     826     121
        Enterobacter aerogenes   hypothetical protein [Klebsiella aerogenes]
gi|844486473|gb|KQ087587.1|:84822-86069  gi|1172898100|ref|WP_080967648.1|
        gi|1172898100|ref|WP_080967648.1|        135     1       30      2.18e-10
        VALTLTGATNRIPNRSPDKALRAAIRESSP   MALTLTGATNRLPNRSPDKALCAAIREPSP   152     63      121
        Enterobacter aerogenes   hypothetical protein [Klebsiella aerogenes]
gi|844486473|gb|KQ087587.1|:84822-86069  gi|1172898100|ref|WP_080967648.1|
        gi|1172898100|ref|WP_080967648.1|        135     27      50      2.18e-10NPPPVALTLTGATNQIPNRSPDKA
        EPSPVALTLTGATNRIPNRSPDKA 73      2       101     Enterobacter aerogenes   hypothetical protein
[Klebsiella aerogenes]
gi|844486473|gb|KQ087587.1|:84822-86069  gi|1172898100|ref|WP_080967648.1|
        gi|1172898100|ref|WP_080967648.1|        135     87      120     1.13e-05
        NPPPVALTLTGATNQIPNRSPDKALCAAIRVFTP       EPSPVALTLTGATNRLPNRSPDKALCAAIREPSP       836     735
        130     Enterobacter aerogenes   hypothetical protein [Klebsiella aerogenes]
gi|844486473|gb|KQ087587.1|:84822-86069  gi|1172898104|ref|WP_080967652.1|
        gi|1172898104|ref|WP_080967652.1|        105     1       105     9.68e-63
        VALTLTGATNRIPNRSPDKARCAAIRESSPVALTLTGATDRIPNRSPDKAPCAAIREPSPVALTLTGATNRLPNRSPDKALCAAIREPSPGGAYAYRG
YQPHTEP
        MALTLTGATNRIPNRSPDKARCAAIRESSPVALTLTGATDRIPNRSPDKAPCAAIREPSPVALTLTGATNRLPNRSPDKALCAAIREPSPGGAYAYRG
YQPHTEP 423     109     523     Enterobacter aerogenes   hypothetical protein [Klebsiella aerogenes]
```



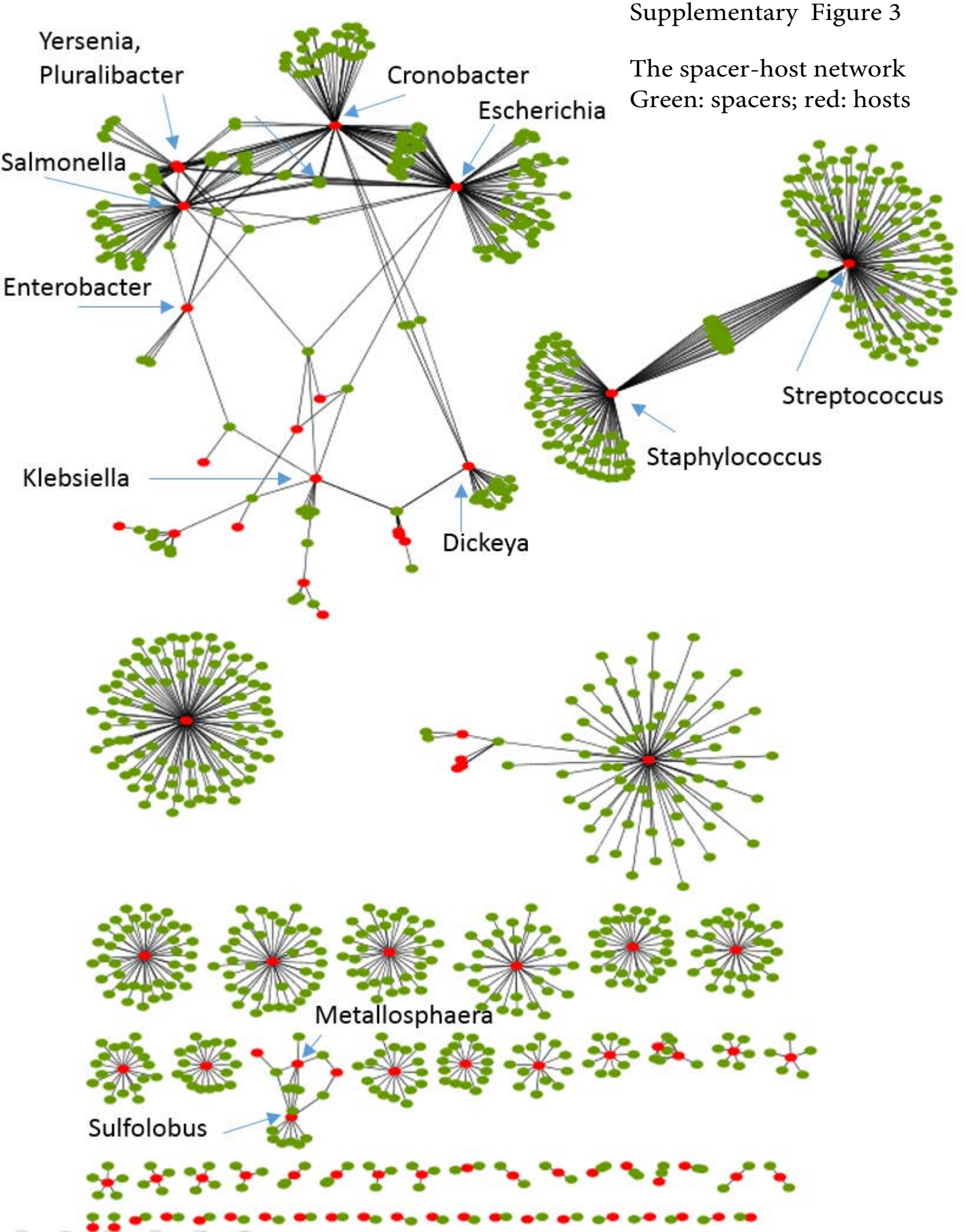
Supplementary Figure 1. The computational pipeline for CRISPR spacer analysis



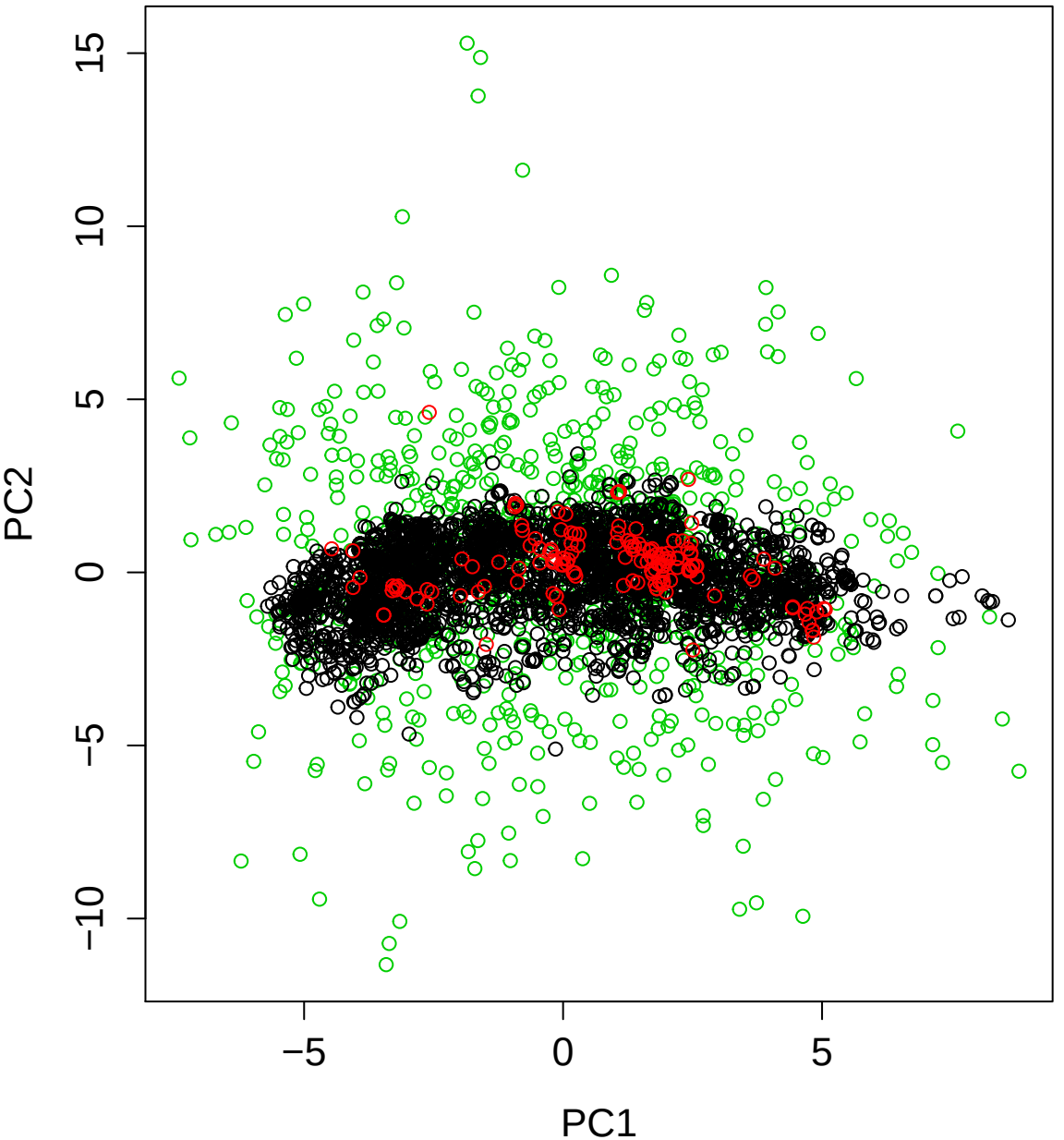
Supplementary Figure 2. Identification of prophage regions with different approaches

Supplementary Figure 3

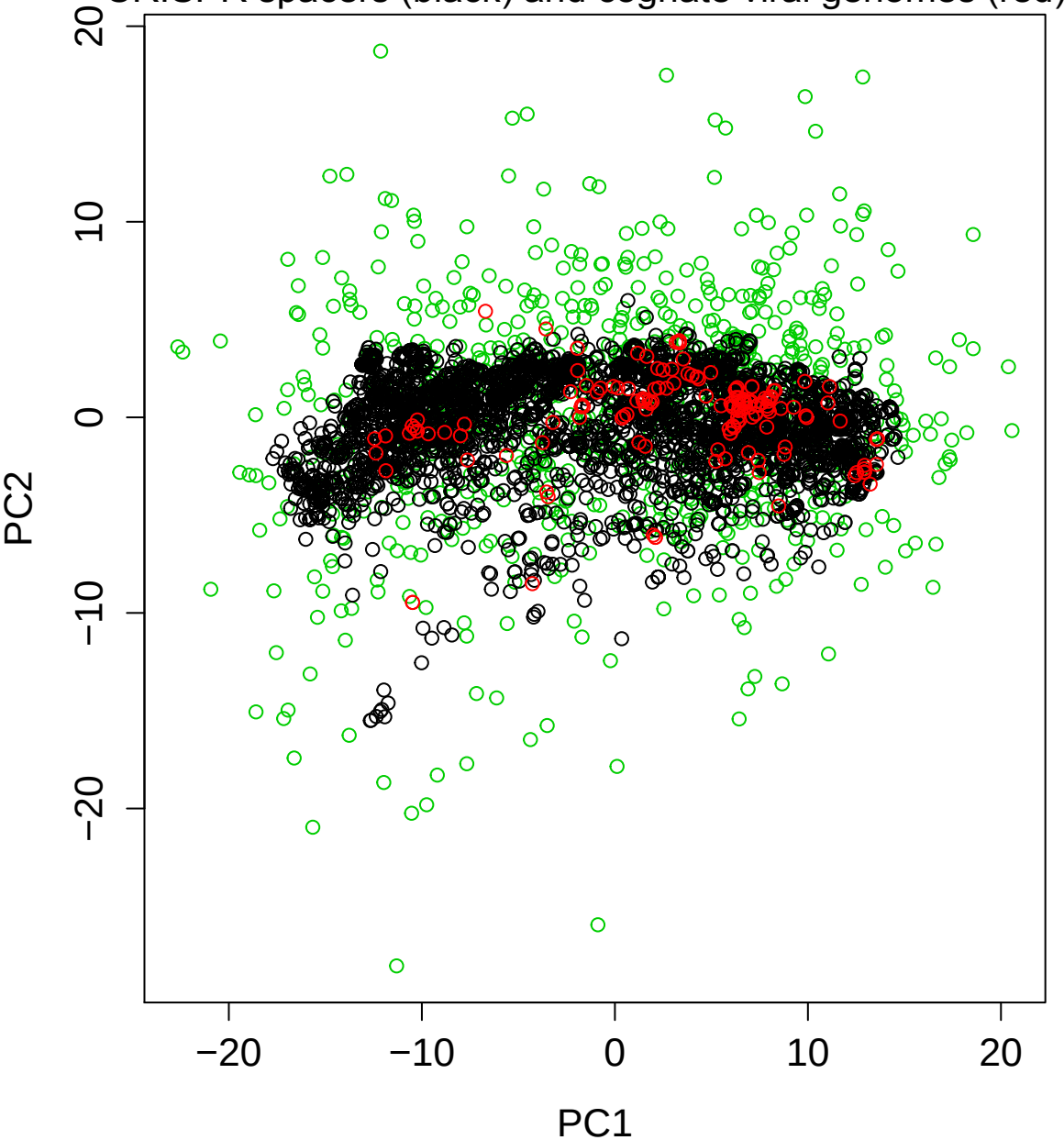
The spacer-host network
Green: spacers; red: hosts



Supplementary figure 4: Principal component analysis of the dinucleotide frequencies in microbial genomes (green), CRISPR spacers (black) and cognate virus genomes (red)



Supplementary figure 5: principal component analysis of tetranucleotide frequencies in microbial genomes (green), CRISPR spacers (black) and cognate viral genomes (red)



Supplementary figure 6 : Defining the threshold for CRISPR spacer identification