Supplementary Information

New classes of self-cleaving ribozymes revealed by comparative genomics analysis

Zasha Weinberg,1* Peter B. Kim,2* Tony H. Chen,3 Sanshu Li,1,2 Kimberly A. Harris,1,2 Christina E. Lünse,2 Ronald R. Breaker1,2,3

1Howard Hughes Medical Institute, 2Department of Molecular, Cellular and Developmental Biology, 3Department of Molecular Biophysics and Biochemistry, Yale University, Box 208103, New Haven, CT 06520-8103, USA.

Correspondence should be addressed to R.R.B. (ronald.breaker@yale.edu)

* These authors contributed equally to this work.
Supplementary Results

Supplementary Data Set 1

Title: printable alignments, genomic contexts and taxonomy of self-cleaving ribozyme candidates

Description: This PDF file contains printable multiple-sequence alignments for all RAGATH motifs, plus pistol and hatchet. It also contains genomic contexts (diagrams of nearby genes and descriptions of conserved domains present in those genes), and the taxonomy of the organisms containing the predicted RNAs, where the organism is known.

Supplementary Data Set 2

Title: machine-readable alignments of self-cleaving ribozyme candidates

Description: Multiple-sequence alignments of the self-cleaving ribozyme candidates are provided in the machine-readable Stockholm format (the format is explained here: http://en.wikipedia.org/wiki/Stockholm_format). These alignments are identical to those in Supplementary Data Set 2, but can be interpreted by a number of computer programs, such as those in Infernal (http://www.ncbi.nlm.nih.gov/pubmed/24008419). These files are therefore useful for readers who wish to perform computational analyses on the alignments, while other readers will likely find the alignments in Supplementary Data Set 1 more useful. The alignments in Data Set 2 are archived as a .zip file.
Supplementary Table 1 | Representative protein domain sequences for each SCRAP (Self-Cleaving Ribozyme Associated Protein). The protein subsequence obtaining the highest score in the final round of a JackHMMER search is listed. This subsequence is assumed to correspond to a conserved domain for the purposes of finding non-coding sequences with a higher chance of containing self-cleaving ribozymes. Some SCRAPs are highly similar (e.g., SCRAP20 and SCRAP21), but did not overlap in JackHMMER searches that were confined to ribozyme-associated proteins. Note that the protein sequence in this table is the highest-scoring after multiple iterations of the JackHMMER search, and are not necessarily the original sequence.

| SCRAP | Highest-scoring protein subsequence |
|-------|-----------------------------------|
| SCRAP1 | NEQLIGACVWLDREQQKSRRAMNSYKAELQARGLAIMEHDHVKNKVFYGDEGSAAITDSMLDILNPDKLKELVGEVYKMKVKEETKTYKFDKFEKAMKAIIFGTYFTETTLEEFEMISKPDDKKQLLKLCKGFEKDKETLISVLPEGETVPFDVELVYIYRIKNGELIKAFLPEEMIDAIIEGIRSKIFVETKTSITL |
| SCRAP2 | EKQRLEASGFARSNGKVLRAVNLIRYNKLTKGSLKDDGVTAEEFLESVFLAEQGYIHRLQIAKESAALSDDSAYAALEAKLTKGGIRLLAGGVEDNMI |
| SCRAP3 | MSEKDDVTVLGLSIPRSIGATFVAGLSCVYSAWIVAGTFTRMESQQAVMAANIEQIKAELVTKNEFERSVQLVNSAIERNREDIRRHELARLEDNQH |
| SCRAP4 | LNQILDKLQMIEHEVSDIKTNMATKQEELVEQNFSTELEDIKANMATKERLEEV |
| SCRAP5 | MALIKCEPPKKSVDQTNQCPNCGRTTADKELAIEQNKSKHHKIALIVIWMILLAVAWGGFYFYQENKRREEQKAEKKAEKKAEKKEKIAEQKQRRQKDFVKTVDLFSDKFSNLSNKISIKATHTWSNAIWKKEKSETNKWTLLKNGKFDDFSDAVNLCISDTVYSNKTAKQYEKFKETYSKWEIKSDDYIYNEKIDICEDTENYCNIAHSLYSLLASPTGNYDDFSTSIITNAENDIKIYAESLSSDMFKLYH |
| SCRAP6 | IKTAEELKEVAHGLSTKPEDICRAODFGRSEVKELIRLDNGRNLNGFDGEPDRGFYSSIGRGLSKYFQVFQFVAFKIWSWETARFYNQHSPFVIDALEENKMLHQVQKELNGLKRAKKDDRDEHRRCREAVIDQAAQSKIGQLEGAEVHDRDMTIMELKAKLYDMM |
| SCRAP7 | MKNGKAPTRQQKKIMKAHGLVPENWLVKNLPSLEVVRVSVLKKVGGKPKTRTISK |
| SCRAP8 | SKAQKAVNKYMSENYDRINLTVPQKEIIKAAHAEAHSESVNGFVNRAI |
AETMERDNAA

SCRAP9
ASKAQQRAVNYMKKNYDRVNLVMPPGRKDAIRAHAEKNGESLNFINRAIDEAMERDDQ

SCRAP10
MGISEAKKRANAKYNEHIIQRLGLLTLEDNEKLGKYLDHIGGQYSELVRNLILEDHMHKGWQELNPPKKKE

SCRAP11
MSELKSGPITKEMWQQIEKEMSDGWNIVNFAYKHEHTTVNRVRVSESCTCLQVYIDGFIRGEWLCSDDGFSDKAPAILPDVGKKTAKYSARFKA
RMTKIGKRGVYKKEWPDCLWLVFYVNPFSKASVLRQYKKTLEGIELVA
AHFVKAEGL

SCRAP12
VVEIQTQVAVIDTNVARIESTIGEVTQVAIIDTNVARIETSIEIQTVQAVIDTNVARIESTIGEVTQVA

SCRAP13
MTMKTMKTVVVFAGAVGMLAACLTVQSSRPVGEALLEILPLMALLLLYVGEAGRLAALQLMGIER
WIDRRENSKSLAWSVLLALGNIKVGEVERPKALGDIRGTYIGM

SCRAP14
MARKRGMQYIPYDYEEAYNKAMEDMHEWFIENLFQHRKKVIYALKEITAGDQFE

SCRAP15
VYFSTADKEISKIGSAERQRQRRAVRLRTEPTEEHLEWVVALYAGVRF
KTYSGLPFTYERIKGRNGQYTKELWIDRRENSKSLAWSVLLALGNIKVGEVERPKALGDIRGTYIGM

SCRAP16
KPKLIQLIHIAKQKLMAMDEYSYRAMLERVTGKTSCKEMSAVEMLKVEAE
MEAKG FKKTSSRNHPSPGKSAVKSRIAYKIRAIESMQKQGLVDRGSEN
ALNAFVRGVVNPIYAKRGMNQLVGLNLQDRASLVLRLFKKWWQR

SCRAP17
MERVQDAETVVEFENEIFMYLIMFCESNAIESEYDILPSQWNAALSYYKHVIKPNDLITPIHTVSNYNINAHDDSALMEYMCYCHNQEISIKGFCLL
SGISRDTIHWSGNSTNTRTAYKIKLDLQGNNISDIAVNILEGEYIKEPESTAAASDI
YKKLKENESSELCQVVDRRPNNPMMKLYPLNKIRSGWNLPGVSKERTSERTALTASELPLKGEVKQIIESEK

SCRAP18
MLRREEYFYEPRTISPAGNIRWFGEIYTAPQLMLCHIEQTQVYIRDNGRMLFI
YELDSKLSEEEKIEAVFTLICIEKTDKGYHRGYRKI

SCRAP19
MNKLSCQLEQLKDNTTKQFENFIVEDLKLALSSYGVVKSREFRKLKNALIT
KYENDELTAIRQALIKCCQNGTDQAICLKYADYFKPETVETVDDGLIEALE
GAGKEAKFDEI

SCRAP20
TLGSRQLTLREEAGFSKRAAEEEAKVYYWLHYEDGRRVPTEPGALRKL
LDAYDVSDE

SCRAP21
TLGSRQLTLREEAGFSKRAAEEEAKVYYWLHYEDGRRVPTEPGALRKL
LDAYDVS

SCRAP22
MEKWIRMAVSVFVLLAAILGLTALTSCGHESGNRKIKNSDSAFVVGIVD
KYCHPESMSSVDEAMLLQQMLMEDNYERVFMGMPQTLKAVHVVM
NKNHGHTPTFTVKAIOQYELSSKRVDNLDPENOQPDPDMVANKO

New Ribozymes
SCRAP23  MNELKGNFKAAYDVVKRAERCVLLRSAANPSHTYAVGDVDSNGLVFG
LVKTDPEAAVREFRTRSFSWPSWKISK

SCRAP24  MEVGRILPTEAVALIVSPQFIRIAMQQGKLPIGTAVQMSSIWTYHISEKLL
ADYSKGDIAERIRGKRA

SCRAP25  MSEKEKQIVKEKLKEVLPNMSDFDKGYLLGKAETMADAADNPKND

SCRAP26  VGNDRYAGVVRLPDAPJAGESYRNAGSHYKVWQYNAADTAHVLENLTSGWVCTAHHIPALYRLHDSVELQWDYSTDGDHE

SCRAP27  KPKTPQEMIEVAKLHIEQGQSINSLAKKYGVSRGRVQNWVRKYQGKEF
DSKRISKTKTKEKLQQAVQLDLYLNNKGSVDIAEKYGILHFLSTFRWIRK
KQGQEL

SCRAP28  MSEKSDKVQPEEIRKGWEEARLQANLREIRGKAKIMIATRKGDTTYRTYTK
K

SCRAP29  IARKVDRLGRIVIPKELRDNMTRITHDDDPLEIYVKDDNIVITKYNPIDD

SCRAP30  GKYQEWLTDPGTRLAEWARDGLTDEQIAARIGITTSTLYDWKKNKYSEFS
EALKRGKEVVDIEVENALLKRALGYDYTEVERSQDGKKISKTVQT
KHIPPDTTAQIFWLKNRRPDRQWDRDKQIEHSGTLE

SCRAP31  MNGAKMLVEVGTVGTGTMGMIAVLAERGRILGLLAVALLI
GLSGVSG
WILGVQMRAKLDRDIWIQGYRAGRTGEMVQKEIRIGIITQK

SCRAP32  IRYRGRFDTGGSIPGEIRYR

SCRAP33  MEERKLVSELWTKGVLKEIAERLNSIVSTVRYYRKL

SCRAP34  MAKETKPKPASTAVATEDNVMEQIKNGNILAEANVKAIEEIQKQKDEK
QKKEAMDMICRAKYLNNKALLELRARRREKKNKEYLTETKNILDEVLG
GKITPIEYKKKCDDLREEFRKKNRESDKQLSEEMQELRESFEGRWQYW
W

SCRAP35  MKIIIEAIAQYEAYGPEYTTLEDAINWWSGPRAWRDFATVLKSGKRPYKQ
HLSEVQFFDDER

SCRAP36  MKRTAWLLAAGLLGMAGAGGALAAASDAGMTISAWAELAVLVLVLAL
GILLYEFGRHKAYRARGYRGFEAGKNAPPDHPSCRHSFALYCEERPGRI
SNWN

SCRAP37  IYECGGTTLSIPKVRDDGNDVFERLSEVVGRNEAIRLCEQFGSNGGGRTYIP
RLSSAMKRYARAHQIREFRLRKHGGASAARVLSVETRMTRNTRNVRWILS

SCRAP38  MKMNKEQFLKTEGFGELESTITAWDDALSRNREDKEVLRTLAWCQQW
EVYYQMAIKHYFGVEYHFRTRTDYEYGLCDESDWLMVVRKVLFQMET
VFSRQVRTRRGRIDGVEYWDZDFVHLNLGKTVIAECTPWLIRVYTEAGT
PLHIFRR

SCRAP39  MKIPKPGQFCTINNVYRAYKAKDGCKGCAFNNLFSCLGIDGKGRFAK
Weinberg, Kim, et. al. and Breaker

New Ribozymes

MDCKYSHIIFKKV

SCRAP40

MRSDAMDALRAALRAFCPTGKETVTVPQLAEALGLTTESQKQLLRRRLILIATRGETEKGVRGWLFWLPIGKEPSRREGESYVRMWRAIRIQSVGWKRODI

ALIARVVTAVGRYWTRLESEGYIAKSQEGSNANLWRTTAKGRDHRQTP

WPVPVDPDYEDERKATAQLCRILLLEDPPKPRTRNRIQEQAVLARFTCT

TGADTPQITEA

SCRAP41

ILNQILDKQMIEHEVSIDTKNMATQKELEEVEKQNFSTELEDIKANMATKR

ELEEVRN

SCRAP42

MSKEEQKPNIQIPKNSDTGYVLYNGQFKTRNFVEIGILALPFGLIFYGIYWH

DLGWDVQSTVAYCIALCAAAALGGAAGHVGDDSLFEFIIIRVRFIRIRSSRIS

KYNMPIKTELEPEYIHHGDRLPKKDKLQMLGTVKGVLEEGNPGFSSDIT

DEELKVFYDDYGFELEKPDALKSAELRAAEKQRAGEEEEEYNNLAISPR

KAAREEIKRKRIDGEAAKKEAERERINRAIKKDLAERVKTAQYYE

KMFMDPNEQNENNAELNPQAVELPVTVPVSNANVQDDIFDFDS

EESITDDEVANEVQADADTELQDGTDTDIAPQIIDDVFDDTEVTLVDV

EDLDDLSVIQDITDSAPQAKQGRQVNVSSDDVSLDDVDLFDFTVSVRT

EERHQKVTSGTMSQKQFPQQTQTNNSDITYLNDISIFDSEE

SCRAP43

MASLAYICRREVSKAGSIDPVAKKLGYARSSLNYLLGLRYPAKSTENIEK

KILATYTSKILCPFTKINIEQKECEDEIVQQNINTGNVLFKLQQFCTCPERL

KNRNEEIKFIRKFEGETLCL

SCRAP44

MDNYNIRFEEIGEQAGQLTMVDDLKAKKAAAIKQKNVVMVESSEDETIE

TIITGSAIDRLGRLGLTLTIETIQMIEKDTDKQYAKAMLYGFVRAIAQAAFERI

SCRAP45

MKEEKLTRLRVRVLVMSYLNLHFLIKEKILSVFLETCLNNNDYSWILKSRIIEQ

GVFQGLSILHDMYEFENYFTFGVSTTLAESAWWRYAFKYYREEPVEV

SCRAP46

MATPESDLRRALLALCPDKTIVCHQQLYEALGLEDAPAKARLRSRIKDL

LHWNDRRKGRDGCTYETHREAMPKRAEGFGMRWAIRAKGPSGWTW

QDIAQITRVYDTVMYVWALDEGFIAQAGRGGTWNSTKGRAEQ

RVAPPPHRPPKDPFEEAGAACAALCLVRLMMEADPNQPHVRKTIVKELGILA

GRFAEAAKEDSHGHDHDPQ

SCRAP47

KPMLDNYQYLLLLEKSLSPNTLEAYLDDAACKLAKLALLEDGIEITVSLEDD

LHGFADAMGLDLJAPRSQARIISGKSLFKLSLDGGYERQDDESLLEAPKIG

RHLPDVLSVEIDSIQINLSTPEGQRKAMITELYSCGLRVSELCTLKME

DLFLDEGIFVSKGKSGQRLVPISDAEHEINLYLTV

SCRAP48

MRKKAFAITISAAICLAALAGCSNKGDDJASISATSETASASSSTSEATDV

TDESLYEFPADALNSEDUMKATISAYADYTQAMSGQGYTYTARFED

GTCHYDGTRTDDANGTDTDDMTLMTYSQIFDAYSAYLNCYQAADANGELLV

GVSEGVEEAVDEAAESEAGQEMAEELSEAESTDSEAEAADGEEAPADD

AVEDDAASSEAEAPAEVEDVASDADEAGDEAAE

SCRAP49

MPKVDAYDPVIVRMLWVAGQLEGLNNKIFTFTKMNDTSVEAENGYRDDL

DGSGVLATCPVRKSDKGGVSVVIIIDWIDLEIPGEAHEAVHAADYMFD

ELGMYQSSFSEHEQYAYLVYGWAFCISKTIKLLKQYDTRRESDDVEDRA
SCRAP50  MEDKVLDTVNGLEYSFEDILVKPLAPIMVTKEYTEQIPTGEKDEEGFN
KYEVKHTHTKEVSEDFAKGIVLISIPTGADSTIKGVDTIVYPKKFAKDFDLFK
DSQLVKPYDVAKVV

SCRAP51  LHKVIDLGAEFALDNFGGIDLYIPVETVSQNRVNRKIIIRDYGEMKNK
SALSRSYQLSYGYTRRIVSSV

SCRAP52  MSTREGYAAALGAGLGMLFGMLFGAIAAGSTWTQGLATLGAIVGLVLGYII
ARMIIEDGRLERELEHQQPDPDEEPAPAMFRDYDTGVRDGYDADAVE
SLQFANLPQRPRKGRKAAAGNMVRAELPERPVMETHQHGAA

SCRAP53  ERGLTLNELSKKAGLSYLSEIERGSKKPSLTDIKAALVNNKQI

SCRAP54  LTQTKEERSLKNRSLSKKNLLETVIMAMAHNTVIKSKHDFGALPAELVA
KYLTIMFGESMTLPLANQTYAAYAVREFVCCQDKGAYLLANSFVKIDSKIESG
ARAFCVLFEMHEVPFGTYCRPSSPWQLAFVAGDITYYIVNISRGSEYQM
GTILNSFGADEDDLKNSRRGILENADMSLPYCGIMQFCVTDDYDIKVE
YAPEDRTIEKAWADQVQR

SCRAP55  MNQINVVTIGKLIEAHRHDGEQKFKFAYVFIAKAYEEQGNDRAANIIERSN
YTGKYGEGQKVLVDIEANGL

SCRAP56  ATFIAJDNTPKLNFSKNSVNTRKIRKVLGTPDNNSERYJAFKQQGIS
VKGKIEDEVVNVKALPIEGDRVQITIIIQSRVVDK

SCRAP57  KTTTFMSDKLEHEKAKEAIKGLGVTLSDFLYAAYVKVL

SCRAP58  EELTALYDNTLNAVKQEVKGVKAIANKKREVIVSVGFKSDGIVPAEFR
YNPLDKGVDTVEVYVVTPECNONGQLESHKQARAQSWDVRNEALNSNE
EIVGYKICRTK

SCRAP59  MDKVAVWNSSSIFRGAYKVYETIIFKNDIARVIENSSKLKNLKVLEYK
KASDRLDSYICISLMLKWQGSDFKHLGARMVTAKTHKYIQDTVFVM
ERQKEAASEIIKILEASVSAKSDTSDK

SCRAP60  MAFPSLNDYKRCIDKNDVKNAGDAAPEASYQPVLNYASTKFKSKLQ
KADVALKSVPDDVLPKLPDNDVIEILKAERSILSVESDEAPAPYQR
KNIATSIPNDFCMVISPQNFKLKLRFFSSLPPPFHEWSYRSKAKASHAVGQY
YLDAEFDYITEAIAKAFAIQGNPILNSEQLFLVCIRGVSPKCTQPSNVDVYW
YTGAITNAICRVLQHGDGDWDMMSFLYTAIPTEYDYLEVILCASKADLFHW
L

SCRAP61  MRRCSRMTEVKDMTLLLSPEDPVRVYKGDIIYNNQYFANMEVDKDIVAQIQ
DAEVKRFRAIPEITHRKYKLERGIAPMPHEETPDYSFDRDLQMCYLYTITI

SCRAP62  SFIKDEVAKKARENLKEERMFMFRRNEVRSFLEGISTLCEVKREAANV
EYLMGKELPDTICVYHDRDITQAVANSHSCRKALLEKQLSLEKVDREV
FKYTESKPWEIASDPCIVIYFSNNQEB

SCRAP63  MNKYYIKEKEDVNYMYEYRGYTISGNDRGYVCISPYGAFSPKAYNTIEDA
KAAIIADDLDDCIAEKYGEQQTREAETTPRIGYIMKALFVEKELYFHLHAAN
DDVDDITYCEDNGEEVLMKNGHVYDIGITGDSPLAAKDVITVMM
YK

SCRAP64 MVKLLNVGSEKAVYSFISFCLGMLAILVLLFRACANLHVVDDDFERIQLQKRTQDLNHKYNYYKATESLLLDSVMEEDDPVLETDCGSNYLDAYRAVKNVLDEKSN

SCRAP65 QKCPLCGGIDPSTGICSVCGIKLV

SCRAP66 MIKIRNNMEYTFKDFGFFKANDVLTWNEDIDAFTMDVEEGNSFRSAMIDERTVEDLRLEGILLEANAEPEDEBIKNTTVFIDSLLLKQYEDDYKEVTQKYKEGKVQPCVKEAETVYFNLTKVLNKIKEELTNE

SCRAP67 MKANLIEMLKAQGFSEVESKSETIQQCNGVLQRDWSRKVEVAHGSYLETYAVRFVNRNSGICHVSYQKDGREYKDRWYDTIGKRTYNAIVETARCAGYAI
**Supplementary Table 2 | Frequency of self-cleaving ribozymes in the vicinity of different genetic elements.** ‘Element name’: refers to entries in the Conserved Domain Database, to twister and hammerhead ribozyme RNAs, or to SCRAP domains. ‘Description’ is a brief description taken from the text in the Conserved Domain Database. The term ‘hypothetical’ denotes putative domains with no known function. ‘Count (element)’: number of occurrences of the element in the RefSeq and environmental sequences databases used in our analysis. ‘type-P1 twister’: fraction of elements that have a type-P1 twister ribozyme within 6 Kb. ‘type-II hammerhead’: fraction of elements with a type-II hammerhead ribozyme within 6 Kb. Only elements with twister or hammerhead fractions of at least 0.01 are listed. The elements are sorted in decreasing order of the higher value of the twister or hammerhead frequency.

| Element name | Description | Count (element) | type-P1 twister | type-II hammerhead |
|--------------|-------------|-----------------|-----------------|-------------------|
| SCRAP18      | Hypothetical (this work) | 134 | 0.470149 | 0.119403 |
| SCRAP28      | Hypothetical (this work) | 97  | 0.463918 | 0.123711 |
| SCRAP31      | Hypothetical (this work) | 161 | 0.322981 | 0.118012 |
| SCRAP32      | Hypothetical (this work) | 160 | 0.28125  | 0.08125  |
| SCRAP63      | Hypothetical (this work) | 118 | 0.084746 | 0.245763 |
| SCRAP35      | Hypothetical (this work) | 170 | 0.041176 | 0.211765 |
| SCRAP38      | Hypothetical (this work) | 261 | 0.206897 | 0.141762 |
| PRK10244     | Hypothetical           | 10  | 0.200000 | 0.200000 |
| SCRAP26      | Hypothetical (this work) | 203 | 0.034483 | 0.17734  |
| SCRAP1       | Hypothetical (this work) | 914 | 0.171772 | 0.080963 |
| SCRAP39      | Hypothetical (this work) | 97  | 0.154639 | 0.120301 |
| SCRAP11      | Hypothetical (this work) | 399 | 0.137845 | 0.120301 |
| pfam09039    | Mu DNA binding, I-gamma subdomain | 705 | 0.107801 | 0.13617  |
| SCRAP23      | Hypothetical (this work) | 130 | 0.076923 | 0.130769 |
| SCRAP36      | Hypothetical (this work) | 39  | 0.128205 | 0.076923 |
| SCRAP45      | Hypothetical (this work) | 212 | 0.117925 | 0.11358  |
| twister, type-P1 | Self-cleaving ribozyme | 1215 | 0.107819 | 0.11358  |
| SCRAP52      | Hypothetical (this work) | 179 | 0.094972 | 0.111732 |
| hammerhead, Self-cleaving ribozyme | 2737 | 0.047132 | 0.109609 |
|                |                                |      |      |
|----------------|--------------------------------|------|------|
| **type-II**    |                                |      |      |
| SCRAP44        | Hypothetical (this work)       | 244  | 0.094262 0.069672 |
| SCRAP7         | Hypothetical (this work)       | 834  | 0.083933 0.059952 |
| SCRAP34        | Hypothetical (this work)       | 429  | 0.083916  |
| SCRAP14        | Hypothetical (this work)       | 546  | 0.080586  |
| pfam11985      | Hypothetical                   | 2013 | 0.077 0.011923 |
| SCRAP64        | Hypothetical (this work)       | 291  | 0.075601  |
| **twister, type-P3** |                     |      |      |
| SCRAP13        | Hypothetical (this work)       | 723  | 0.067773 0.042877 |
| COG4734        | Anti-restriction protein        | 50   | 0.06 |
| pfam02914      | Mu transposase                 | 705  | 0.014184 0.051064 |
| pfam10805      | Hypothetical                   | 317  | 0.050473  |
| pfam06252      | Hypothetical                   | 2603 | 0.043027 0.049558 |
| SCRAP66        | Hypothetical (this work)       | 474  | 0.048523  |
| SCRAP59        | Hypothetical (this work)       | 111  | 0.045045  |
| pfam09299      | Mu transposase, C-terminal     | 4139 | 0.024402 0.043247 |
| pfam08765      | Mor (Middle-operon regulator) transcription activator | 3270 | 0.042508 0.030581 |
| pfam02316      | Mu DNA-binding domain (MuA transposase and repressor protein CI) | 907  | 0.040794  |
| SCRAP50        | Hypothetical (this work)       | 649  | 0.035439  |
| SCRAP19        | Hypothetical (this work)       | 1186 | 0.035413  |
| pfam11145      | Hypothetical                   | 29   | 0.034483  |
| SCRAP22        | Hypothetical (this work)       | 1372 | 0.031341 0.000729 |
| SCRAP25        | Hypothetical (this work)       | 1031 | 0.031038  |
| COG2842        | Uncharacterized ATPase / putative transposase | 5539 | 0.027081 0.030511 |
| SCRAP16        | Hypothetical (this work)       | 7001 | 0.024854 0.02214 |
| SCRAP41        | Hypothetical (this work)       | 3464 | 0.023961 0.004619 |
| PRK00635       | Excinuclease ABC, subunit A    | 126  | 0.02381  |
| SCRAP61        | Hypothetical (this work)       | 909  | 0.006601 0.023102 |
| PRK05784       | Phosphoribosylamine-glycine ligase | 44   | 0.022727  |
| pfam07352      | Mu Gam-like protein (Bacteriophage Mu Gam protein protects linear dsDNA from exonucleases) | 1770 | 0.018079 0.022034 |
| HYPOTHETICAL (ork work) | 0.018849 | 0.005899 |
|------------------------|----------|----------|
| DNA topoisomerase I    | 110      | 0.018182 |
| SpoVT / ArbB-like (transcriptional regulation) | 3234 | 0.017934 | 0.005257 |
| Hypothetical           | 176      | 0.017045 |
| Hypothetical           | 2624     | 0.016768 |
| Glutathione synthetase (eukaryotic subtype) | 61 | 0.016393 |
| Homocysteine methyltransferase | 124 | 0.016129 |
| DNA-directed RNA polymerase, mitochondrial type | 189 | 0.015873 |
| d-alanyl-d-alanine carboxypeptidase | 189 | 0.015873 |
| Ribosomal RNA methyltransferase N | 207 | 0.014493 |
| Hypothetical           | 415      | 0.014458 |
| Enoyl-(acyl carrier protein) reductase | 141 | 0.014184 |
| Phage/conjugal plasmid zinc-finger protein, type C-4 | 714 | 0.014006 |
| DNA-binding protein     | 219      | 0.013699 |
| Hypothetical           | 1849     | 0.013521 |
| Endoplasmic reticulum oxidoreductin 1 | 152 | 0.013158 |
| Toxin-co-regulated pilus synthesis protein Q | 76 | 0.013158 |
| Hypothetical           | 156      | 0.012821 |
| Phage portal protein, A118 family | 4923 | 0.012594 | 0.001422 |
| Self-cleaving ribozyme | 345      | 0.002899 |
| Hypothetical           | 87       | 0.011494 |
| Sugar fermentation stimulation protein A | 271 | 0.01107 |
| Hypothetical           | 6510     | 0.010753 |
| BetR domain            | 750      | 0.010667 |
| Paired-Box domain      | 96       | 0.010417 |
| Hypothetical           | 1250     | 0.0104  |
| Lower collar protein   | 1061     | 0.010368 |
| Lipid-4'-kinase        | 687      | 0.010189 |
| SURF4 family           | 197      | 0.010152 |
**Supplementary Table 3 | Conserved elements selected to define regions enriched for self-cleaving ribozymes.** Conserved elements whose name starts “pfam” refer to entries in the Conserved Domain Database. Elements are listed in order beginning with the strongest candidate elements.

| Conserved element name                                  | Number of kilobases considered in 5’ direction (relative to the element) | Number of kilobases considered in 3’ direction |
|--------------------------------------------------------|--------------------------------------------------------------------------|-------------------------------------------------|
| twister ribozyme, Type-P1                              | 6                                                                        | 6                                               |
| twister ribozyme, Type-P3                              | 6                                                                        | 6                                               |
| twister ribozyme, Type-P5                              | 6                                                                        | 6                                               |
| hammerhead ribozyme, type-I                            | 6                                                                        | 6                                               |
| hammerhead ribozyme, type-II                           | 6                                                                        | 6                                               |
| hammerhead ribozyme, type-III                          | 6                                                                        | 6                                               |
| SCRAP18                                                | 6                                                                        | 6                                               |
| SCRAP35                                                | 6                                                                        | 6                                               |
| SCRAP38                                                | 6                                                                        | 4                                               |
| SCRAP26                                                | 4                                                                        | 4                                               |
| pfam09039                                              | 4                                                                        | 6                                               |
| pfam02914                                              | 4                                                                        | 6                                               |
| SCRAP11                                                | 3                                                                        | 3                                               |
| SCRAP1                                                 | 3                                                                        | 2                                               |
| SCRAP45                                                | 3                                                                        | 3                                               |
| SCRAP44                                                | 2                                                                        | 2                                               |
| SCRAP7                                                 | 3                                                                        | 3                                               |
| SCRAP14                                                | 2                                                                        | 0                                               |
| SCRAP22                                                | 4                                                                        | 0                                               |
| pfam09299                                              | 3.5                                                                      | 4                                               |
| pfam06252                                              | 6                                                                        | 3                                               |
| SCRAP61                                                | 7                                                                        | 0                                               |
Supplementary Table 4 | Genetic elements involved in the detection of confirmed self-cleaving ribozyme. IGRs nearby to certain genetic elements were used because of a putative relationship between those genetic elements and self-cleaving ribozymes. The observation that ribozymes were in fact discovered in the vicinity of these element classes can be viewed as evidence supporting the proposed association, at least for those element classes. For each self-cleaving ribozyme class, certain IGRs were used in the original purely computation prediction of a conserved motif, and these IGRs were present because they were close to some of the genetic elements. In the table below we list which genetic element classes were actually responsible for the inclusion of those IGRs, for each of the 5 confirmed self-cleaving ribozymes (including diverged versions of previously known ribozyme classes). In cases where more than one genetic element class was responsible for the same IGR, both classes are listed. “RAGATH” is “RNAs Associated with Genes Associated With Twister and Hammerhead”. The 15 motifs detected in the first round of searching are named RAGATH-1 through -15.

| Ribozyme                        | Genetic element(s)                                      |
|---------------------------------|--------------------------------------------------------|
| Variant hammerhead (RAGATH-1)   | Hammerhead, type-I                                      |
|                                 | Twister, Type-P1                                       |
| Variant HDV (RAGATH-2)          | SCRAP22                                                  |
| Twister sister (RAGATH-3)       | pfam06252                                               |
|                                 | pfam09299                                               |
| Pistol                          | Hammerhead, type-I                                      |
|                                 | Hammerhead, type-II                                     |
|                                 | Twister, Type-P5                                        |
| Hatchet                         | SCRAP22                                                  |
|                                 | SCRAP45                                                  |
|                                 | Hammerhead, type-III                                     |
|                                 | Twister, Type-P1                                        |
Supplementary Table 5 | Properties of the 12 motifs that did not self-cleave under our experimental conditions. The consensus primary and secondary structures of these motifs appear in Supplementary Figure 1. Domains elucidated for this study that were not present in the Conserved Domain Database were named SCRAPs (Self-Cleaving Ribozyme Associated Proteins). “RAGATH” is “RNAs Associated with Genes Associated With Twister and Hammerhead”. The 15 motifs detected in the first round of searching are named RAGATH-1 through -15.

| Motif name | Notes |
|------------|-------|
| RAGATH-4   | RAGATH-4 is a 3-stem junction that is above average in size for a bacterial ncRNA (~130 nt), and has multiple regions of sequence conservation distributed throughout its predicted secondary structure. Representatives are frequently flanking by twister and/or hammerhead ribozymes. The structure of RAGATH-4 is larger than most self-cleaving ribozymes that do not depend on a cofactor, with the exception of the Neurospora VS ribozyme that is of comparable complexity. Rho-independent intrinsic transcription terminators are usually predicted immediately downstream of RAGATH-4 RNAs. |
| RAGATH-5   | RAGATH-5 is a small structure with a predicted pseudoknot. The 3’ end of the motif contains the conserved sequence CCA, which is also present on the 3’ ends of mature tRNAs. However, there are no other data to suggest a relationship with tRNAs, so the similarity might be a coincidence. RAGATH-5 does not have a clearly discernible association with flanking genes, partially because the metagenomic sequences in which it is found are often short contigs. However, the most common nearby genetic elements with an annotated function are type-II hammerhead, SCRAPs 8 and 9 (which are very common, and therefore not necessarily indicative of ribozyme function), and methyltransferases. RAGATH-5 RNAs are also frequently located nearby to other RAGATH-5 RNAs, which is a property similar to known self-cleaving ribozymes in bacteria. |
| RAGATH-6   and -7 | A striking feature of RAGATH-6 and RAGATH-7 is that they are usually located near each other, surrounding a twister ribozyme. In most cases, a RAGATH-6 RNA is found immediately 5’ to a twister ribozyme, and then immediately 3’ to the twister ribozyme is a RAGATH-7 RNA. However, in some cases the RAGATH-6 or the RAGATH-7 RNA is not detected. These data suggest one of two hypotheses. First, the arrangement of a twister ribozyme flanked by RAGATH-6 and -7 RNAs could represent a single large |


RNA that uses the twister ribozyme to achieve a more complex function such as allostERIC control (see ref #1 at the end of this file). In this model, the cases where either RAGATH-6 or -7 is missing could be explained by alternate structures used by a subset of RNAs or by events such as insertion of selfish genetic elements. However, a second hypothesis is that RAGATH-6 and -7 represent RNAs that function as independent structures, but are often encoded adjacent to twister ribozymes.

RAGATH-8 is a moderately large (average of 274 nt) and relatively complex structure. Some representatives of this motif appear to be missing some regions of the RNA, or to have nucleotides that fail to align well in certain regions. As a result, there are fewer highly conserved nucleotides in the RNA when all these representatives are aligned. RAGATH-8 RNAs appear in a variety of gene contexts, only some of which involve nearby elements that are characteristic of validated self-cleaving ribozymes.

RAGATH-9 is a somewhat large (average of 164 nt) RNA that appears in similar gene contexts, wherein many of the genes are in the same orientation as the RNA. Most examples are nearby to SCRAP22 and SCRAP34 or SCRAP45 domains, as well as tRNAs or tmRNAs and a predicted sigma factor 70 gene.

The most common gene context of RAGATH-10 RNAs involves the appearance of a type-III hammerhead hundreds of base pairs downstream. The intervening space usually contains a hypothetical gene that is often classified as SCRAP62. An engA ribosome-associated GTPase gene is often found upstream of the RAGATH-10 RNA. The RNA itself is a 3-stem junction, although the enclosing stem is not highly supported by covariation.

RAGATH-11 RNAs are frequently associated with twister ribozymes. However, the available RAGATH-11 RNAs have rather similar sequences and might be closely related. Therefore, it is possible that the apparent ribozyme association does not represent a true functional association.

RAGATH-12 RNAs are often found downstream of SCRAP64 domains and near twister ribozymes. They consist of two hairpins, and a conserved sequence 5′ to these hairpins. No structure was detected in this 5′ sequence, but the lack of mutations makes it impossible to find covariation.

RNAs are generally found in the order Legionellales of γ-Proteobacteria. They are often found nearby to hammerhead ribozymes.

Most RAGATH-14 RNAs are followed by either a type-III hammerhead ribozyme or a
predicted pseudo-tRNA. A twister ribozyme is often located farther downstream.

| RAGATH-15 | RAGATH-15 RNAs are often found upstream of SCRAP64 domains, with twister ribozymes often located 3-4 Kb away. Although the motif is quite large (average of ~450 nt), the structure consists largely of a series of hairpins, with only one multistem junction present. |
Supplementary Figure 1 | Additional candidate RNA motifs, RAGATH-4 to RAGATH-15. “RAGATH” is “RNAs Associated with Genes Associated With Twister and Hammerhead”. The 15 motifs detected in the first round of searching are named RAGATH-1 through -15. Annotations are as described in Fig. 1a. Note that a small number of homologs for RAGATH-8 do not fit the alignment well, and when these are removed, many nucleotides are conserved at least 97% (colored red). Other properties of these motifs are described in Supplementary Table 5.
Supplementary Figure 2 See next page for legend.
**Supplementary Figure 2 | In vitro transcription of self-cleaving ribozyme candidates.** (Top) Sequences and secondary-structure models for a hammerhead representative, an HDV representative (see Supplementary Fig. 4), and two twister sister representatives (TS-1 and TS-2). Red nucleotides are highly (≥97%) conserved among the variant RNAs identified in this study. For each construct, disruptive mutations were designed by altering either one or more highly-conserved nucleotides (circled: M1, M2, M3, M4, and M7), or structured regions (boxed: M5 and M8). Compensatory mutations (boxed: M6 and M9) carry nucleotide sequences that are different than the WT RNAs, but that restore base pairing. (Bottom) Denaturing (8 M urea) 20% PAGE separations of RNA transcription and ribozyme cleavage products generated by *in vitro* transcription of DNA to generate internally $^{32}$P-labeled RNA. FL denotes bands corresponding to full-length transcripts. Clv denotes predicted product bands due to self-cleaving ribozyme activity. **Note:** The mutant full-length HDV-1 RNA (sharp band) and its anomalous extended products (slower running smeared band) do not cleave during transcription and therefore are nicely visible in their expected locations. Anomalous extension of RNA transcripts to make these spurious longer products is a common problem of T7 RNA polymerase. The WT HDV-1 RNA lane has this same pattern (sharp band and the slower running smear), but these RNAs have efficiently cleaved due to the action of the ribozyme, and therefore are shorter on their 5’ ends. Thus, the full-length WT RNA transcripts are actually not robustly seen on the gel because nearly all are cleaved during the transcription reaction.
Supplementary Figure 3 | Cleavage site analysis of a bimolecular hammerhead ribozyme HH-1. (a) Sequence and secondary structure model of the bimolecular hammerhead construct HH-1. Annotations are as described in the legend to Fig. 2a. (b) Mass spectrum analysis of the products of the bimolecular HH-1 cleavage reaction. Enzyme and substrate RNAs were incubated under standard reaction conditions (see supplementary methods). Annotations are as described in the legend to Fig. 2d.
Supplementary Figure 4 | A bimolecular construct of a variant HDV ribozyme representative (RAGATH-2) and its cleavage site. (a) Sequence and secondary structure of a bimolecular HDV construct from a metagenomic sequence (accession SRS024075_LANL_scaffold_3320, nucleotides 35795 to 35731, on the reverse-complement strand). This ribozyme has a predicted E-loop structure, as do 46% of RAGATH-2 HDV variants (Fig. 1a). Red nucleotides are highly (≥97%) conserved among the variant RNAs identified in this study. The substrate strand is identified by the gray box, and the nucleotides outside the box constitutes the enzyme strand. The cleavage site expected based on prior work on HDV ribozymes is also confirmed in our experiments and is indicated by the arrow labeled “Clv”. M1 designates an enzyme strand construct wherein C64 has been mutated to a U residue. (b) PAGE separation of the ribozyme reaction products. Annotations are the same as in Fig. 2c, where S is the 5'-32P-labeled 19 nt RNA substrate. The labeled substrate (~5 nM) was incubated with ~100 nM enzyme in the absence or presence of 10 mM Mg²⁺ at 23°C for 20 minutes. Additional details are described in Online Methods.
Supplementary Figure 5 | Ribozyme activity and cleavage site mapping of a bimolecular construct of TS-2. (a) Sequence and secondary structure model of a bimolecular construct derived from TS-2 (metagenomic sequence origin). Other annotations are as described in the legend to Fig. 2a. (b) PAGE separation of ribozyme cleavage products. S designates the 5’ 32P-labeled 25-nucleotide RNA substrate (trace amounts), and C14 identifies the 5’ 32P-labeled fragment band produced by incubation with excess unlabeled ribozyme for 30 min (Rxn lane). Other annotations are as described in the legend to Fig. 2c. See Online Methods for experimental details.
Supplementary Figure 6 | Internal phosphoester transfer mechanism for RNA cleavage. All validated classes of natural self-cleaving ribozymes exploit this general cleavage mechanism. The new chemical groups resulting from this reaction include a 2’,3’-cyclic phosphate on the 5’ cleavage fragment and a 5’ hydroxyl group on the 3’ cleavage fragment. Many natural and artificial RNA-cleaving enzymes, whether they are composed of RNA, DNA, or protein, employ multiple catalytic strategies in combination to accelerate this chemical transformation. The figure is adapted from a previous publication.⁸
Supplementary Figure 7 | Twister sister ribozyme activity is abolished with a deoxyribonucleotide at the site of cleavage. (a) Sequence and secondary structure of TS-1 bimolecular complex wherein C13 is replaced with a 2’-dC nucleotide. Other annotations are as describe for Fig. 2a. (b) Ribozyme cleavage assays using the bimolecular construct TS-1. C and dC refer to substrates that contain C or 2’-dC, respectively, at position 13.
Supplementary Figure 8 | Kinetic analyses of the bimolecular constructs TS-3 (environmental sequence, accession SRS013951_WUGC_scaffold_60933, nucleotides 2461 to 2542) and TS-4 (environmental sequence, accession SRS014979_C3268593, nucleotides 976 to 1053). (a) Sequence and secondary structure model of a bimolecular construct derived from TS-3. Annotations are as described in the legend to Fig. 2a. (b) PAGE autoradiogram of a representative time course for cleavage of the bimolecular TS-3 construct depicted in a. Annotations are as described in the legend to Fig. 2b. (c) Dependence of the rate constant for TS-3 on pH. Note that the maximum rate constant of ~2 min⁻¹ is achieved at a suboptimal Mg²⁺ concentration. (d) Dependence of the rate constant for TS-3 on Mg²⁺ concentration. Note that the maximum rate constant of ~2 min⁻¹ is achieved at a suboptimal pH. Under optimal conditions for pH and Mg²⁺ concentration, the rate constant for TS-3 is predicted to be greater than 100 min⁻¹, which is too fast to measure using our assay method. (e) Sequence and secondary structure model of a bimolecular construct derived from TS-4. Annotations are as described in the legend to Fig. 2a. (f) PAGE autoradiogram of a representative time course for cleavage of the bimolecular TS-4 construct depicted in e. Annotations are as described in Fig. 2b. Note that the reaction is already complete after an 18 second incubation at a low Mg²⁺ concentration, suggesting that TS-4 is also promoting RNA cleavage with a rate constant that is much faster than the αγ speed limit.
Supplementary Figure 9 | Twister sister cleavage activity with various metal cations. (a) 32P-labeled substrate RNA for the TS-1 bimolecular construct (Fig. 2a) was incubated with excess enzyme RNA in the absence (-) or presence of 1 mM of the indicated divalent metal ion for 60 min under otherwise standard reaction conditions. (b) TS-1 bimolecular was incubated in the absence (-) or presence of 1 M of the indicated monovalent metal ion for 60 min under otherwise standard reaction conditions. Possible contaminating divalent metal ions were chelated by the addition of 30 mM EDTA. “No reaction” samples (NR) were mixed with five volumes of stop solution (90% formamide, 50 mM EDTA, 0.05% xylene cyanol, 0.05% bromophenol blue) prior to addition of monovalent ions. (c) Bimolecular TS-1 ribozyme assay in the absence (0 mM) or presence of 1 mM or 10 mM Co(NH3)63+ incubated without Mg2+ at 23°C for 0 to 48 h.
Supplementary Figure 10 | Observed rate constants for bimolecular constructs of pistol and hatchet ribozyme representatives. (a) (Top) Sequence and secondary structure of a pistol ribozyme construct from Lysinibacillus sphaericus (RefSeq accession NC_010382.1, nucleotides 2010928-2011021) used for a bimolecular ribozyme cleavage assay. Red nucleotides denote highly conserved nucleotides according to the consensus sequence model for this ribozyme class. Other annotations are described in the legend to Fig. 2. (Bottom) A time course assay of pistol ribozyme cleavage wherein the substrate RNA strand is $^{32}$P-labeled. Reactions were incubated at 23°C, aliquots were terminated at the times indicated, and the products were separated by denaturing (8 M urea) 20% PAGE. The amounts of substrate and 5′ cleavage product were...
quantified and plotted to estimate $k_{obs}$ (see **Online Methods**). Under standard buffer conditions (30 mM HEPES, pH 7.5, 100 mM KCl) with 1 mM MgCl$_2$, the $k_{obs}$ value is 0.9 min$^{-1}$. (b) (Top) Sequence and secondary structure of a hatchet ribozyme construct from a metagenomic sequence (SRS017191_Baylor_scaffold_14517/1281-1081) used for a bimolecular ribozyme cleavage assay. A time course assay of hatchet ribozyme cleavage wherein the substrate RNA strand is $^{32}$P-labeled. Reactions were performed as described in a, and the $k_{obs}$ is 0.6 min$^{-1}$ under standard buffer conditions in the presence of 10 mM MgCl$_2$. 
Supplementary Figure 11 | Activity of a bimolecular twister sister ribozyme. The data shown are identical to those in Figure 2, except that full-length versions of the gel images are included. The figure legends are also as in Figure 2.
Supplementary Figure 12 | Structure and activity of pistol self-cleaving ribozymes. The data shown are identical to those in Figure 3, except that full-length versions of the gel images are included. The figure legends are also as in Figure 3.
Supplementary Figure 13 | Structure and activity of hatchet self-cleaving ribozymes. The data shown are identical to those in Figure 4, except that full-length versions of the gel images are included. The figure legends are also as in Figure 4.
Supplementary References

1. Lee, E. R., Baker, J. L., Weinberg, Z., Sudarsan, N. & Breaker, R. R. An allosteric self-splicing ribozyme triggered by a bacterial second messenger. *Science* **329**, 845-848 (2010).