Electronic Supplementary Information

# RNA diversification by a self-reproducing ribozyme revealed by deep sequencing and kinetic modelling

Cyrille Jeancolas,[ab*] Yoshiya J. Matsubara,[c*] Mykhailo Vyborniy,[a] Camille N. Lambert,[a] Alex Blokhuis,[d] Thomas Alline, Andrew Griffiths,[a] Sandeep Ameta,[c] Sandeep Krishna[c] and Philippe Nghe [a]

[a.] Laboratoire de Biochimie, UMR CNRS-ESPCI 8231, Chimie Biologie Innovation, PSL University, ESPCI Paris, 10 rue Vauquelin, 75005 Paris, France

[b.] Laboratoire d'Anthropologie Sociale, UMR 7130, Collège de France, 52 rue du Cardinal Lemoine, 75005 Paris, France

[c.] Simons Centre for the Study of Living Machines, National Centre for Biological Sciences, Bellary Road, Bangalore 560 065, Karnataka, India

[d.] Groningen Institute for Evolutionary Life Sciences, University of Groningen, 9747 AG Groningen, The Netherlands

* These authors contributed equally.

Email: philippe.nghe@espci.psl.eu

## This file includes :

# Mechanistic details

*Mechanistic details for Repeated Transfer (RT)*

See **Fig. 2d** for a graphic representation of the reaction scheme. The succession of trans-esterifications involved in RT has been coined "R2F2" in the context of the *Azoarcus* ribozyme self-assembly.[1] The two steps of the RT mechanism are reminiscent of the reverse and forward reactions of the second step of self-splicing in group I introns.[1,2] In the first step, the IGS of the ribozyme binds to the 5'-most tag of **RNA1** and the 3'-OH of the ribozyme's 3'-guanosine attacks the phosphodiester bond immediately downstream of the 5'-most tag, the resulting transesterification reaction leading to the elongation of the ribozyme and cleavage of the substrate. Then, the elongated ribozyme can bind to the 3'-most tag of another **RNA1**, initiating the second step of the RT mechanism. The 3'-OH of **RNA1** attacks the phosphodiester bond between the ribozyme and the mobile unit, and the resulting transesterification reaction adds the mobile unit to the 3'-end of **RNA1**. Elongation by the RT reaction scheme is different from the elongations shown by Cech and coworkers.[3–5] Indeed, the RT reaction does not oligomerize nor circularize the ribozyme itself (in *cis*),[3] but other RNA fragments (in *trans*) which enables multiple turnovers. The RT reaction also neither requires sequence complementarity for the added nucleotides nor a guanosine leaving group,[4] which enables indefinite recombinations. Finally, the RT reaction does not require strong sequence complementarity (such as 11 nt) between the ribozyme and the substrate, or a guanosine at the oligomer 3'-end, and does not require the oligomerization to happen at the 3'-end of the ribozyme itself.[5] This means the RT reaction offers a very flexible way to oligomerize RNA sequences.

*Mechanistic details for Terminal Strand Attack (TSA)*

Based on prior mechanistic knowledge,[1] we deduce that the 3'-OH of one substrate strand attacks a sterically proximal phosphodiester near the 5'-end of the other substrate strand, the transesterification reaction leading to formation of a hairpin structure. The ribozyme can also bind to the substrate at the 3'-tag of the newly formed hairpin, and circularizes it using the same mechanism. These trans-esterifications are reminiscent of the forward reaction of the second step of self-splicing in group I introns, and have been named "tF2" in the context of the *Azoarcus* ribozyme self-assembly.[1] To our knowledge, no circularization of an RNA other than the ribozyme itself,[6] has been shown through this mechanism.

# Figures S1 to S9

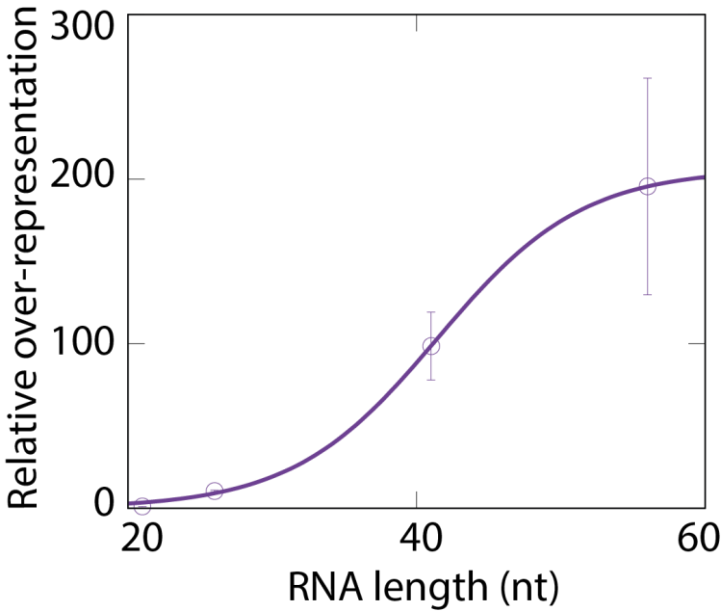



**Fig. S1.** Calibration curve for sequencing data. The sequencing procedure biases the number of sequences toward longer ones (over the size range in this study). Here, four RNAs (corresponding to **RNA1** stem + 0, 1, 4 and 7 mobile units corresponding to $S_0$, $S_1$, $S_4$ and $S_7$) were mixed in equimolar proportions and sequenced. The x-axis corresponds to the length of the RNA, and the y-axis corresponds to the relative over-representation of reads compared to the smallest RNA sequenced, that is $\frac{number\ of\ reads\ of\ S_n}{number\ of\ reads\ of\ S_0}$. The error bars correspond to the standard errors extracted from triplicates. The calibration curve (solid line) is $y = \frac{a}{b+\exp(-c(n-21)/5)}$, with $n$ being the length of the RNA, y the relative over-representation. The fitted parameters are a = 3.5 $\pm$ 1.3, b = 0.017 $\pm$ 0.006 and c = 1.0 $\pm$ 0.1.

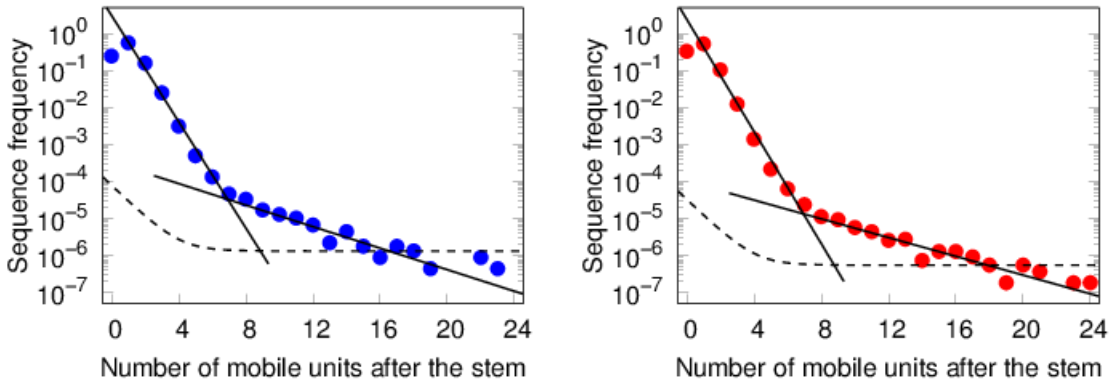



**Fig. S2.** (a) In blue. This graph shows all the product sequences $S_n$ detected in the entire calibrated sequencing data and their relative proportions, in the **RNA1** processing after 45min of reaction. The data points up to $S_7$ are the same as in **Fig. 2c**. The maximum sequence length detected is $S_{23}$. There are two exponential decay regimes with the sequence length, i.e., the number of units (regression lines for each are shown as solid black lines); the former ($S_1$ ~ $S_7$) is predicted by the kinetic model (**Fig. 2c**), while the latter ($S_8$ ~ $S_{23}$) is not taken into account in the model. The dashed line indicates the calibrated limit of detection (Material and Methods and **Fig. S1**). (b) In red. The same plot displaying the **RNA3** processing (**Fig. 4b**).

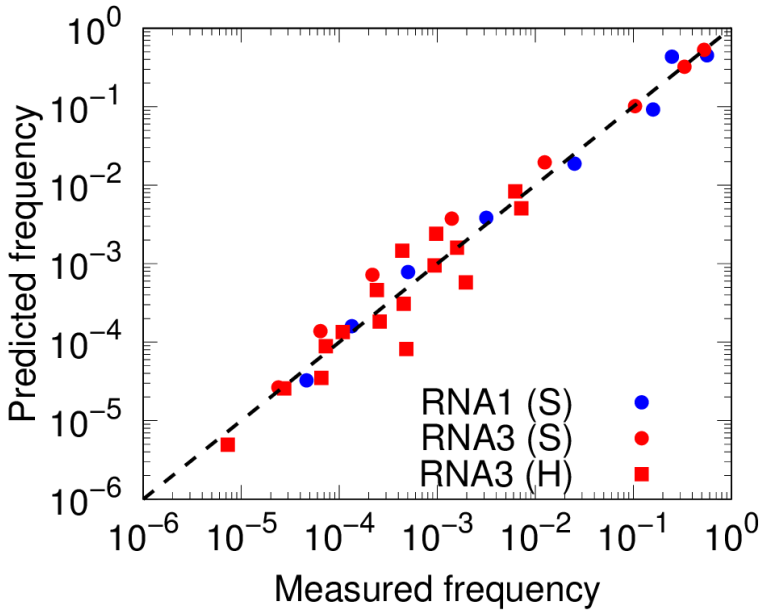



**Fig. S3.** The scatter plot for the measured (x-axis) and predicted (y-axis) frequencies of RNA sequences (the same data set as in **Fig. 2c** or **Fig. 4b**). Circle and square points denote the data set for $S_n$ ($0 \leq n \leq 7$) and $H_{m,n}$ ($0 \leq m, n \leq 3$), respectively. The colors (blue or red) represent the data set in **RNA1** or **RNA3** processing, respectively. The dashed line is for y=x. The correlation coefficient $r$ is calculated as $R^2 = 0.988$ for **RNA1** (blue), and $R^2 = 0.940$ for **RNA3** (red) and the p-value is calculated as $6.0 \times 10^{-7}$ and $6.5 \times 10^{-15}$ for **RNA1** and **RNA3** respectively.

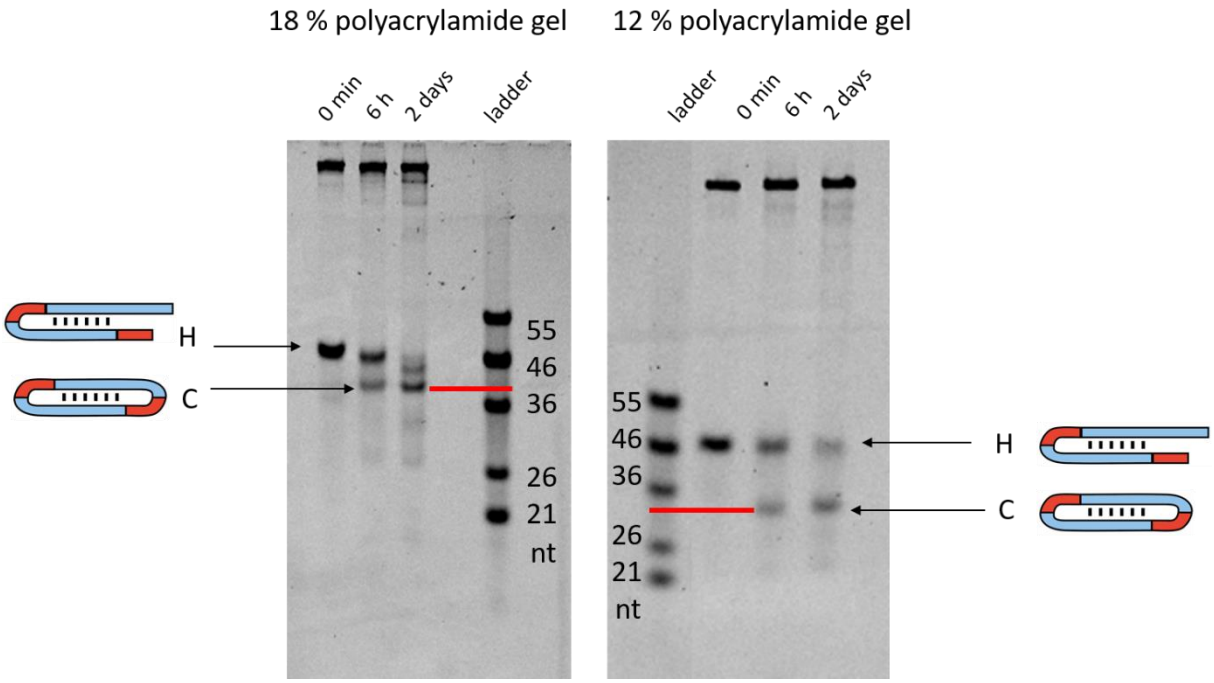



**Fig. S4.** Demonstration of the cyclic nature of certain RNA products. The hairpin RNA (H) has the sequence of both **RNA2**'s and **RNA3**'s intermediate product when processed by *Azoarcus* (confirmed with sequencing). The sample was run on both 18% and 12% polyacrylamide gels. The band corresponding to the circular RNA (C) product migrates slower than the 36 nt reference band (see ladder) in the 18 % gel and faster than the same band in the 12% gel which is the hallmark of circular RNA.[7,8]

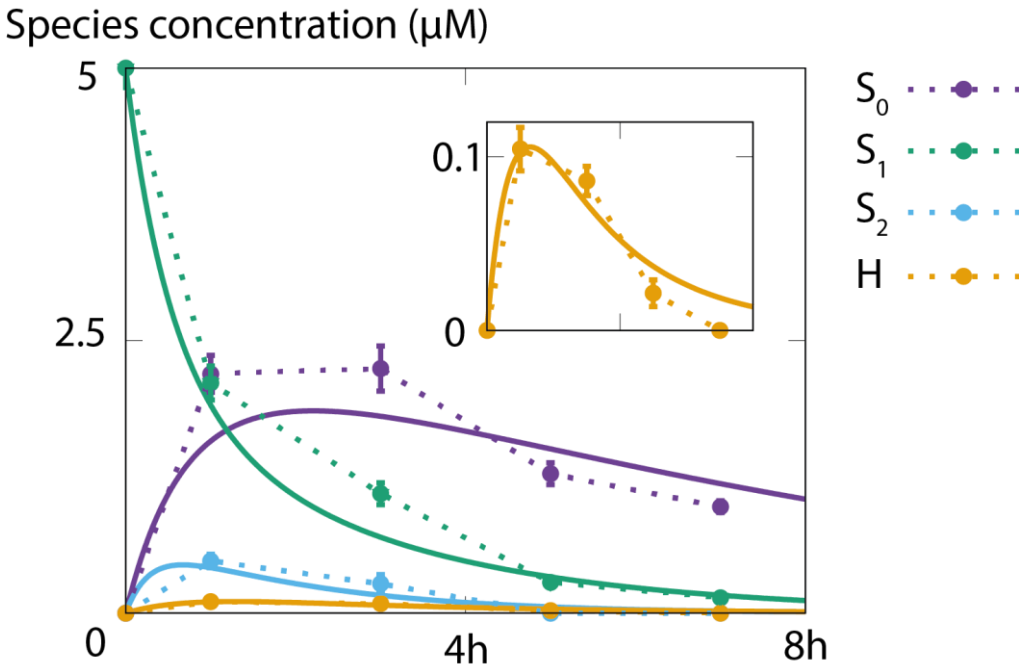
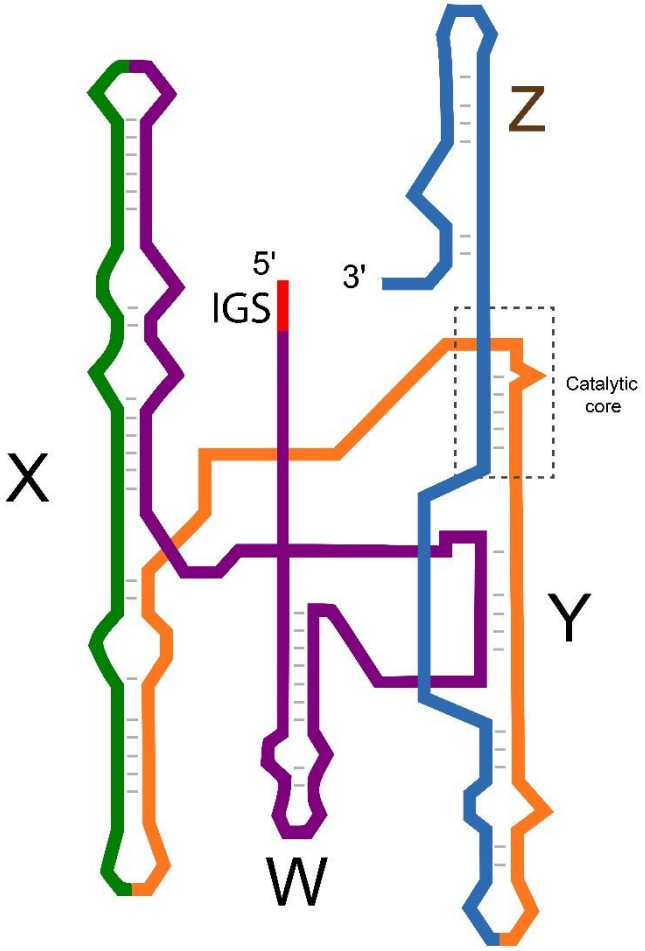



**Fig. S5.** Graph showing the evolution of the concentrations of **RNA3** ($S_1$), product without a mobile unit ($S_0$), product with two mobile units ($S_2$), and hairpins (H), when reacting with the covalent *Azoarcus* ribozyme. The experimental values extracted from the band intensities are represented by dots (connected by a dotted line) and the concentrations predicted by the model are represented by solid lines. The error bars represent standard deviations from triplicates. The conditions for the reaction are the same as in **Fig. 2b** and **3b**. The inset depicts H species concentration variation at lower values.




WXYZ sequence:

GUGCCUUGCGCCGGGAAACCACGCAAGGAAUGGUGUCAAAUUCGGCGAAACCUAAGCG
CCCGCCCGGGCGUAUGGCAACGCCGAGCCAAGCUUCGGCGCCUGCGCCGAUGAAGGUG
UAGAGACUAGACGGCACCCACCUAAGGCAAACGCUAUGGUGAAGGCAUAGUCCAGGGA
GUGGCGAAAGUCACACAAACCGG

**Fig. S6** Scheme of the full length *Azoarcus* ribozyme (WXYZ) secondary structure.[9] The W, X, Y and Z parts are depicted in purple, green, orange, and blue respectively. The IGS (the internal guide sequence which corresponds to the 5' GUG) is depicted in red.

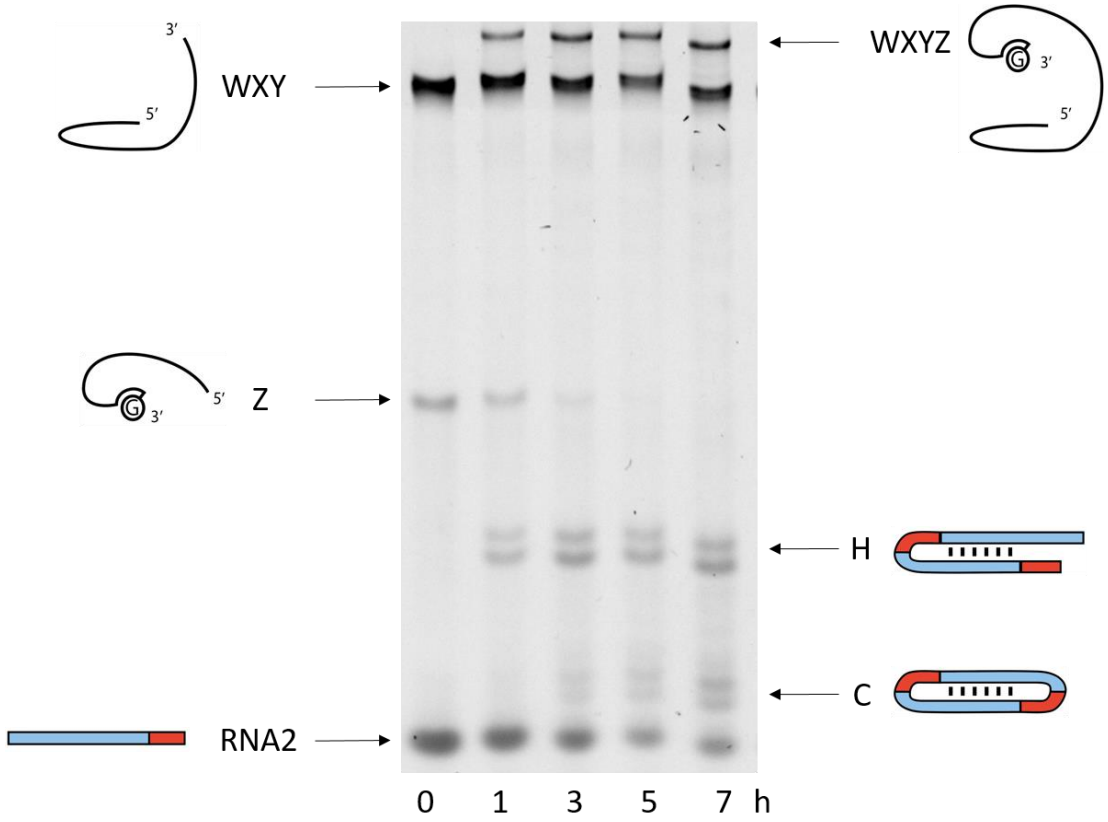



**Fig. S7. RNA2** processing and *Azoarcus* self-reproduction from WXY and Z on a 12 % polyacrylamide gel. The double bands for **RNA2** products account for two possible sites for transesterification, as observed in sequencing data. The experimental conditions are the same as in **Fig. 5b**.

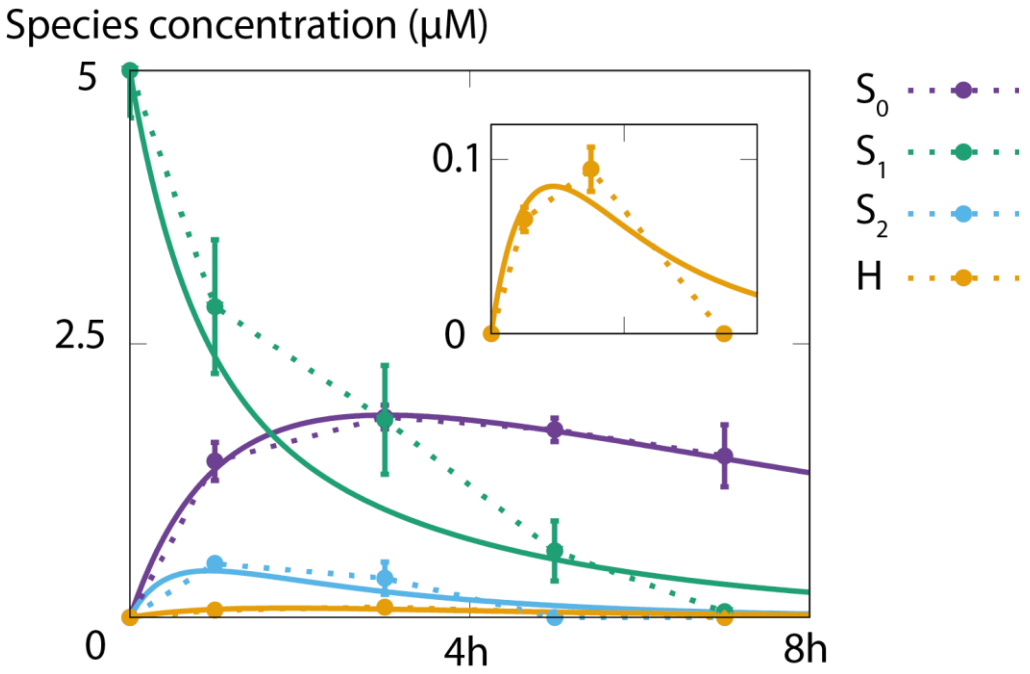



**Fig. S8.** Time course of **RNA3** processing seeded with WXY and Z instead of WXYZ. The time course of the product is similar to with **RNA3** + WXYZ (see Fig. **S5**), with newly fitted parameters, which account for the inhibitory effects of free Z strand and the lower efficiency of the WXY:Z non-covalent complex (see Model for RT+TSA mechanism in Material and Methods). The experimental values extracted from the band intensities are represented by dots (connected by a dotted line) and the concentrations predicted by the model are represented by solid lines. The error bars represent standard deviations from triplicates. The dynamics predicted by the combined kinetic model match well with the experimental data. The inset depicts H species concentration variation at lower values.

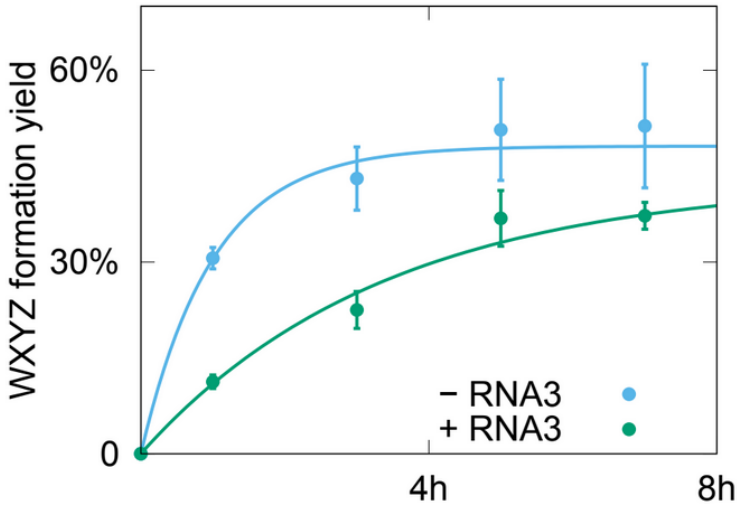



**Fig. S9.** Time course of WXYZ production with and without **RNA3**. WXYZ still forms despite the presence of 10 times more of the competitor (RNA3) with catalytic rates reduced by around 75%. The initial conditions are: [WXY] = [Z] = 0.5 µM and [RNA3] = 5 µM. The dots represent PAGE data and the solid lines represent the theoretical fitting (see Model for WXYZ self-reproduction in Materials and Methods).

# Materials and methods

*Materials*

All chemicals were purchased from Merck (unless specified otherwise). For all the reactions, water was used from ThermoFisher Scientific (UltraPureTM DNAse/RNase free) or from a MilliQ water purifier system (Millipore). RNA concentrations were measured on a NanoDrop-1000 UV-spectrophotometer (Peqlab). Denaturing polyacrylamide gels were prepared using gel stock solution from Roth and run in 1 × TBE (Tris-Borate ethylenediaminetetraacetic acid (EDTA)), prepared from 10 × TBE from Roth). All analyses were performed using 12% to 18% denaturing polyacrylamide gels containing 8.3 M urea and run for at least 2–3 h at constant power of 24 W. Gels were stained with 1× SYBR Gold (ThermoFisher Scientific). Gel analysis and calculation of conversions were carried out with ImageJ software (https://imagej.nih.gov/ij/). The extraction of relevant RNA concentrations from band intensities were carried out by calibration to band intensities of same-length RNA of the same length with known concentrations from 1 to 5 µM. All DNA oligonucleotides were obtained from IDT DNA technologies (https://eu.idtdna.com) and are described in **Table S1**.

**Table S1 DNA primers and templates for PCR and *in vitro* transcription used in the study**

| Oligo name | Sequence (5' to 3') | Description |
|---|---|---|
| Primer1 | CTGCAGAATTCTAATACGACTCACTATAGTGCCTAGCGCCGGGAAACCACGCTAGGGATGG | Forward primer to generate dsDNA template, by PCR, adding a T7 promoter (underlined), for transcription of the WXYZ ribozyme or WXY fragment with GUG as IGS |
| Primer 2 | ATGTGCCTTAGGTGGGTGC | Reverse primer to generate dsDNA template to produce WXY with CAU tag |
| Primer 3 | TAATACGACTCACTATAGGCATCGCTATGGTGAAGGCATAG | Forward primer to generate dsDNA template to produce Z |
| Primer 4 | CCGGTTTGTGTGACTTTCGCC | Reverse primer to generate dsDNA template to produce Z and WXYZ |
| Primer 5 | GTG ACT GGA GTT CAG ACG TGT GCT CTT CCG ATC TTTT TTT TTT TTT TTT VN | Primer for cDNA synthesis and to add Read2 (library preparation) |

| | | |
|---|---|---|
| Primer 6 | ACA CTC TTT CCC TAC ACG ACG CTC TTC CGA TCT -rGrGrG | Template Switching Oligo, to add Read1 (library preparation) |
| Primer 7 | AATGATACGGCGACCACCGAACAC TCTTTCCCTACACGACGCTCTTCCG ATCT | P5-Read1 (library preparation) |
| Primer 8 | CAAGCAGAAGACGGCATACGAGAT -index- GTGACTGGAGTTCAGACGTGTGCTC TT | P7-index-Read2 (library preparation) |
| WXYZ template DNA | GTGCCTTGCGCCGGGAAACCACGCAAGGA ATGGTGTCAAATTCGGCGAAACCTAAGCGC CCGCCCGGGCGTATGGCAACGCCGAGCCA AGCTTCGGCGCCTGCGCCGATGAAGGTGTA GAGACTAGACGGCACCCACCTAAGGCAAA CGCTATGGTGAAGGCATAGTCCAGGGAGT GGCGAAAGTCACACAAACCGG | Plus strand of the dsDNA template used to generate WXYZ, WXY and Z dsDNA templates, by PCR, for *in vitro* transcription . |

*RNA preparation*

The *Azoarcus* ribozyme (WXYZ) and WXY RNA fragment were prepared by *in vitro* transcription. The dsDNA templates were produced using standard PCR reactions (see below). For PCRs ~18 pg of plasmid bearing dsWXYZ was mixed with 1x PCR buffer (ThermoFisher Scientific), 0.75 µM of each forward and reverse primer (see **Table S1**), 0.2 mM of each dNTP, 0.02 U/µL Hot Start Phusion polymerase (ThermoFisher Scientific, product no.: F-549-L) and thermocycled as follows: step 1: 98 °C / 30 s, step 2: 98 °C / 10 s, step 3: 57 °C / 30 s, step 4: 72 °C / 30 s with 24 additional cycles from step 2 to 4 and final extension at 72 °C / 3 min. The purity of the dsDNA was checked on a 2% agarose gel (stained with GelRed™, run under standard electrophoresis conditions; 1 × Tris-acetate-EDTA (TAE), 110 V, 40 min). After PCR, amplified dsDNA templates were ethanol precipitated (2.5 x volume, -80 °C for 30 min, centrifuged for 30 min at 11 000 rcf), pellets were washed with 70% ethanol, dissolved in water and used directly for *in vitro* transcription. *In vitro* transcription was performed using HiScribe™ T7 High Yield RNA Synthesis Kit (New England BioLabs) at 50 µL scale, the RNAs were purified by a phenol-chlorophorm extraction (Acid-Phenol:Chloroform, pH 4.5, Invitrogen) followed by an ethanol precipitation. The RNAs were resuspended in RNase free water, mixed with an equivalent volume of gel loading buffer (70% formamide, 130 mM EDTA, 0.1% xylene cyanol, 0.1% bromophenol blue) and purified on 12% denaturing polyacrylamide gels (urea 50%) using standard electrophoresis conditions (1 × TBE buffer, run at 24 W for 2 - 3 h). Transcript bands were excised, crushed and eluted in 500 µL 0.3 M Na-Acetate overnight at 26 °C. The eluted solution was micro-filtered with 0.2 µm Minisart Syringe Filters, ethanol precipitated, washed with 70 % ethanol, dissolved in water and

concentrations measured with Nanodrop (ND-ONE-W ThermoFisher Scientific). Z fragment and other RNA oligonucleotides mentioned in this study were from Integrated DNA Technologies and used without further purification.

*Trans-esterification RNA reactions*

For the trans-esterification reactions, *Azoarcus* ribozyme (0.5 µM), or WXY and Z (0.5 µM each), were mixed in water with the RNA substrate (5 µM each). To fold the RNA, the mixture was heated at 80 °C for 3 min and gradually cooled down to 20 °C (at a rate of 0.1 °C/s). Then 60 mM of $MgCl_2$ was added and the reaction was incubated at 44 °C. Self-reproduction experiments of **WXYZ** followed the previously established protocols, including the use of an equimolar mixture (0.5 µM) of two RNA derived by fragmenting *Azoarcus* – **WXY** and **Z**.[10,11] In this study, the ribozyme contains 5'-GUG as an IGS and all RNA substrates 3'-CAU as a tag. Experimental data were acquired by polyacrylamide gel electrophoresis (PAGE) in triplicate and species distributions measured by next-generation sequencing by adapting the Smart-Seq protocol. The parameters of the kinetic model of the reactions were obtained after fitting to PAGE data only.

*RNA library preparation and Next Generation Sequencing*

The reaction mixtures for sequencing were prepared as described above and stopped with one volume of the gel loading buffer. After denaturing gel extraction and ethanol precipitation with glycogen carrier (Merck), the RNAs were dissolved in water. The RNA libraries were prepared according to the SMARTseq procedure.[12] For the polyadenylation step, ~ 200 ng of RNA were mixed with 2 mM of ATP, 60 U/µL of Poly(A) Polymerase (from Yeast, ThermoFisher Scientific), 1 x corresponding reaction buffer, and incubated for 10 min at 16 °C. The solutions were then cooled on ice for 1 min, mixed with 17 µM of polyT oligonucleotide (**see Table S1**), incubated for 3 min at 72 °C and cooled on ice again for 1 min. For the cDNA synthesis step, the solutions were mixed with 0.5 mM of each dNTP, 5 mM of DTT (1,4-dithiothreitol), 5 µM of the template switching oligonucleotide (**see Table S1**), 24 U/µL of SuperScript III (ThermoFisher Scientific), 1 x of the corresponding buffer, and incubated for 60 min at 55 °C and for 15 min at 70 °C. For the PCR step, 40 % of the previous solutions were mixed with 0.2 mM of dNTP, 0.5 µM of indexed forward and reverse primers (**see table**), 0.02 U/µL of Hot Start Phusion polymerase and 1 x of the corresponding buffer and thermocycled as follows: step 1: 98 °C / 30 s, step 2: 98 °C / 15 s, step 3: 72 °C / 45 s with 12 additional cycles from step 2 to 3 and final extension at 72 °C / 3 min. Libraries were barcoded using P5 primers containing 6 nt sequences from the NEXTflex series (Illumina). dsDNAs were extracted and eluted with a PCR Clean-up kit (Macherey Nagel). Length profiles were assessed with capillary electrophoresis (Agilent 2200 TapeStation, using high sensitivity D1000 ScreenTape®, Product No.: 5067-5584) and non-specific amplifications (shorter than the library minimal size and longer than 700 nucleotides) were cut off using AMPure XP Beads.

The concentration of the dsDNA library was measured using Nanodrop, Qubit (dsDNA HS Assay Kit) and qPCR with the NEBNext Library Quant Kit for Illumina (New England BioLabs). The indexed libraries were pooled and ~ 8 ng were sequenced with 30% PhiX using the Illumina NextSeq 550 system in 2*150 High Output mode at the Genotyping and Sequencing Core Facility, ICM Paris (iGenSeq, Institut du cerveau et de la moelle épinière).

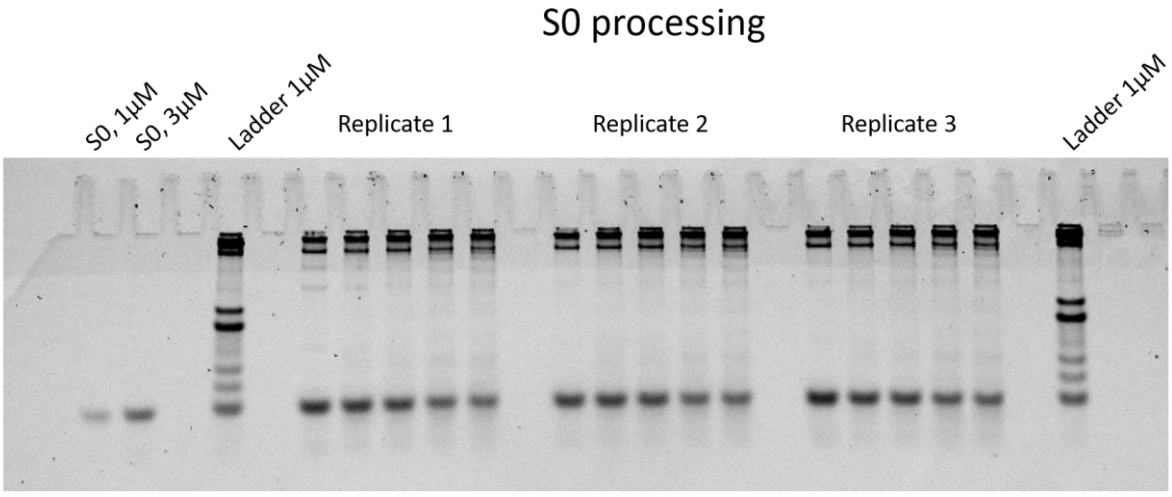
*Sequencing data processing*

FASTQC files were processed by a custom Python script. The files were first cleaned and the reads trimmed after the polyA tail to discard ligation products formed during library preparation. All sequences were sorted according to their sequence and counted to generate the distributions shown in **Fig. 2c** and **Fig. 4b**.

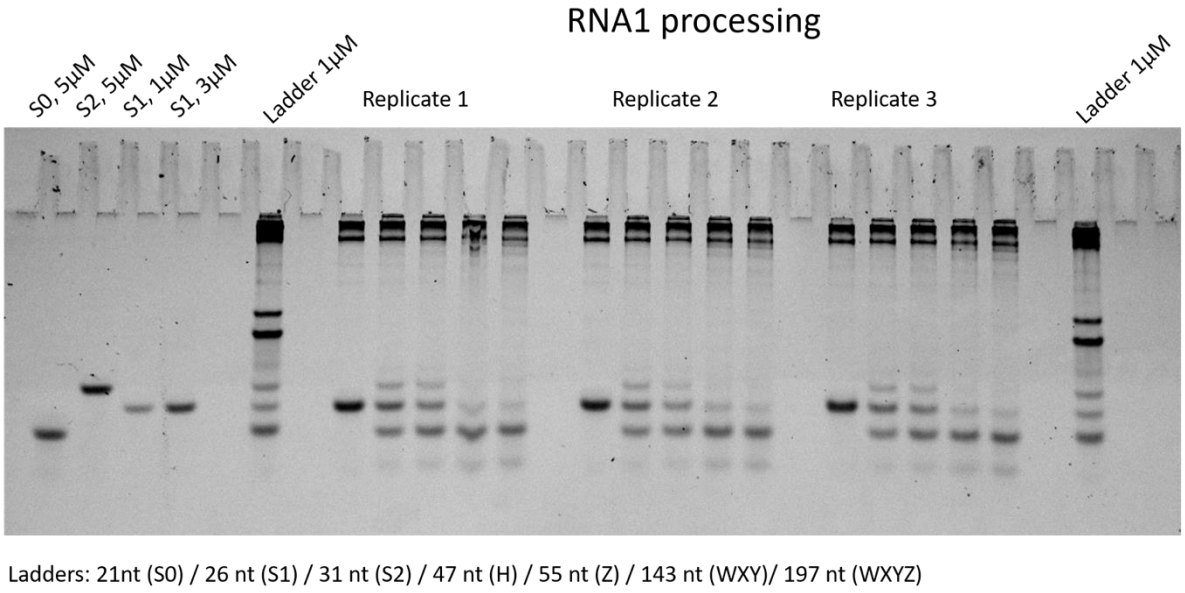
The limits of detection of RNA sequencing (dashed lines in **Fig. 2c** and **Fig. 4b**) are defined as the RNA species frequency $f$ above which there is more than 95% probability to detect at least one sequencing read. These limits are computed assuming a Poisson statistic of species capture from the RNA mixture, resulting in $f = \frac{-\ln(0.05)}{N}$, where $N$ is the number of reads used for the analysis. For **RNA1** and **RNA3** sequencing, we find $f = 2.43 \times 10^{-5}$ and $f = 9.11 \times 10^{-6}$ respectively. For consistency with the calibrated sequencing data, the dashed lines in **Fig. 2c** and **Fig. 4b** represent these limits after calibration and normalization (see **Fig. S1**).

## Gel images used for the study

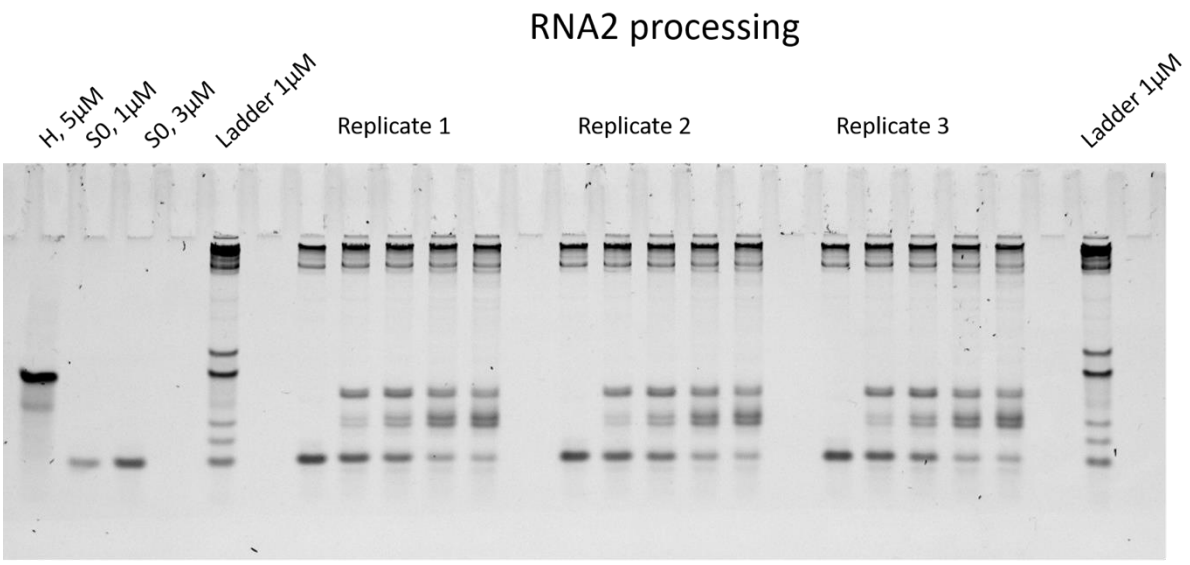
All the poly-acrylamide gels are 18% acrylamide unless stated otherwise. They all show the same reaction in triplicate. The time points for all replicated reactions displayed on the following gels (Replicate 1, Replicate 2, Replicate 3) are from left to right: 0h, 1h, 3h, 5h, 7h. All the reactions are set with 60 mM of $M_gCl_2$ at 44 °C. The bands at the left of the first ladder correspond to known fragments at different concentration used to calibrate the size-dependent ratio between band intensity and RNA concentration.



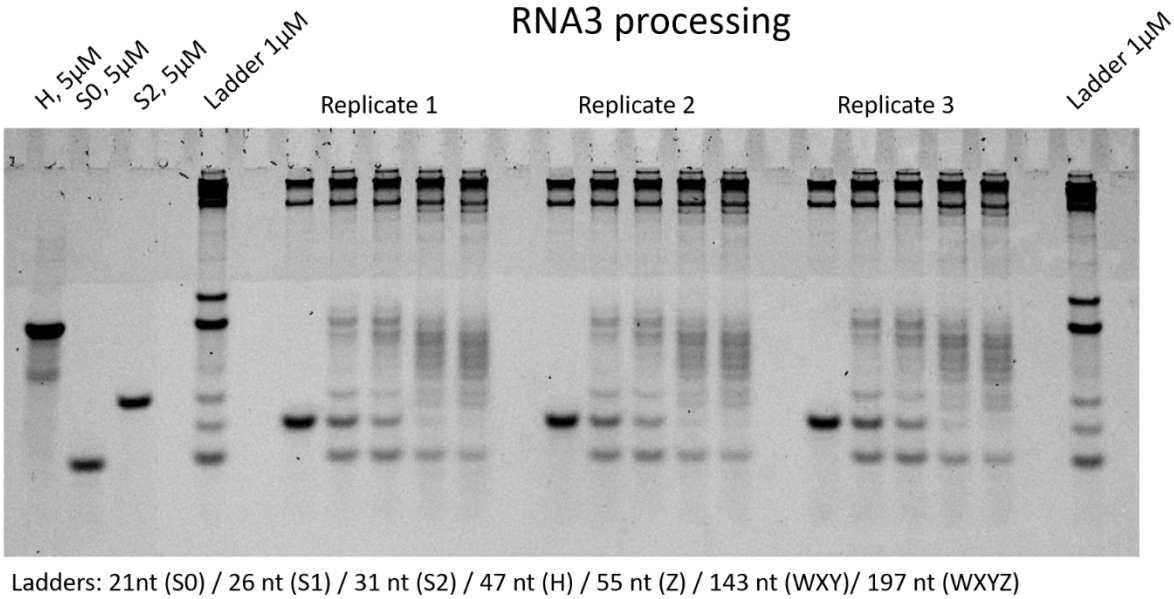

**Gel 1.** Stability assay of the stem $S_0$ (5 µM) of RNA1 in contact with *Azoarcus* (0.5µM). The time points for Replicate 1, Replicate 2 and Replicate 3 are from left to right: 0h, 1h, 3h, 5h, 7h.



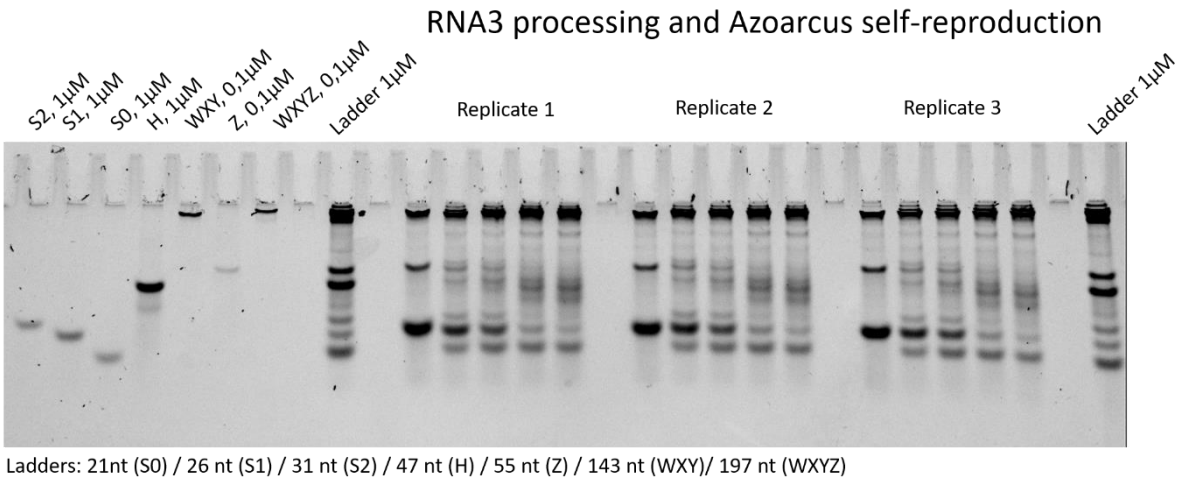

**Gel 2**. RNA1 (5 µM) processed by *Azoarcus* ribozyme (0.5 µM). The time points for Replicate 1, Replicate 2 and Replicate 3 are from left to right: 0h, 1h, 3h, 5h, 7h.



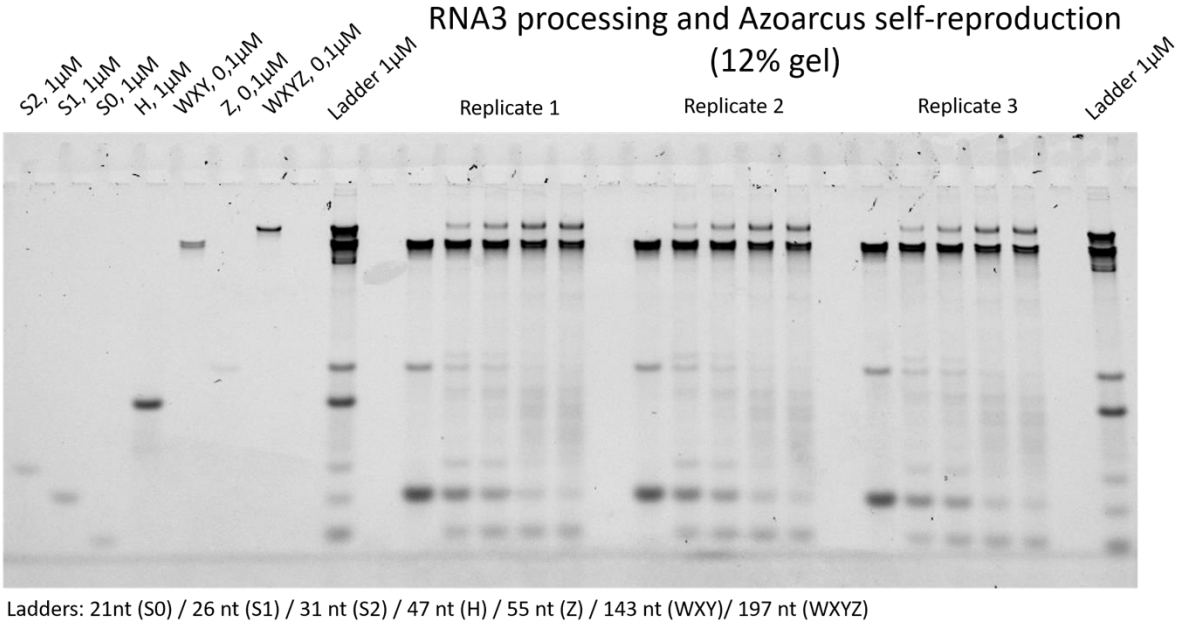

**Gel 3.** RNA2 (5 µM) processed by *Azoarcus* ribozyme (0.5 µM). The time points for Replicate 1, Replicate 2 and Replicate 3 are from left to right: 0h, 1h, 3h, 5h, 7h.



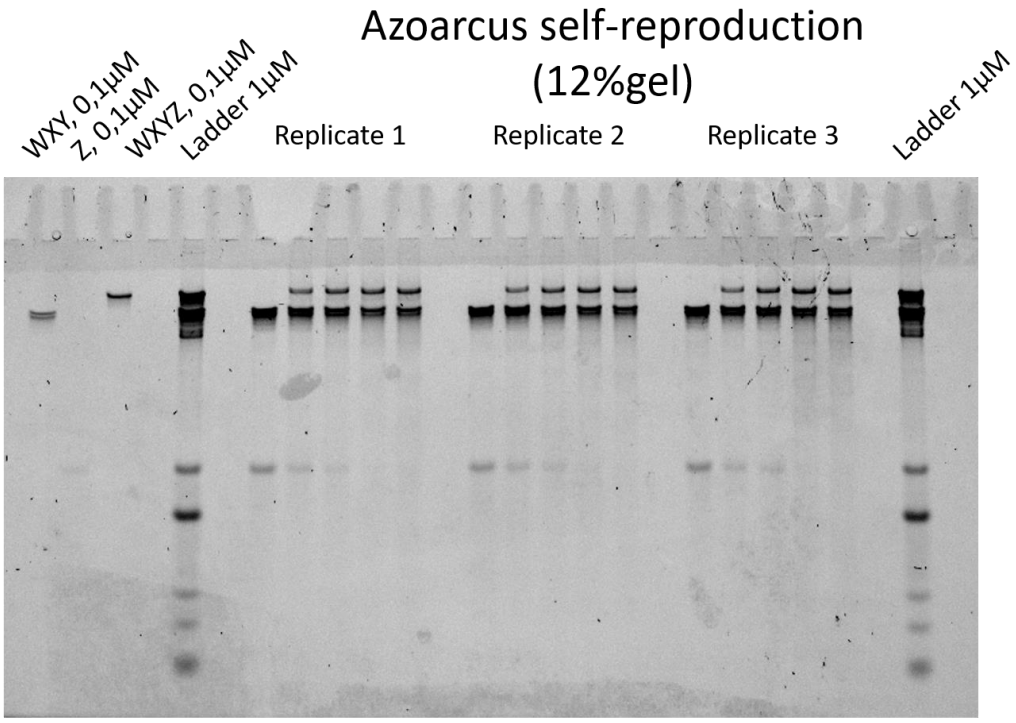

**Gel 4.** RNA3 (5 µM) processed by *Azoarcus* ribozyme (0.5 µM). The time points for Replicate 1, Replicate 2 and Replicate 3 are from left to right: 0h, 1h, 3h, 5h, 7h.



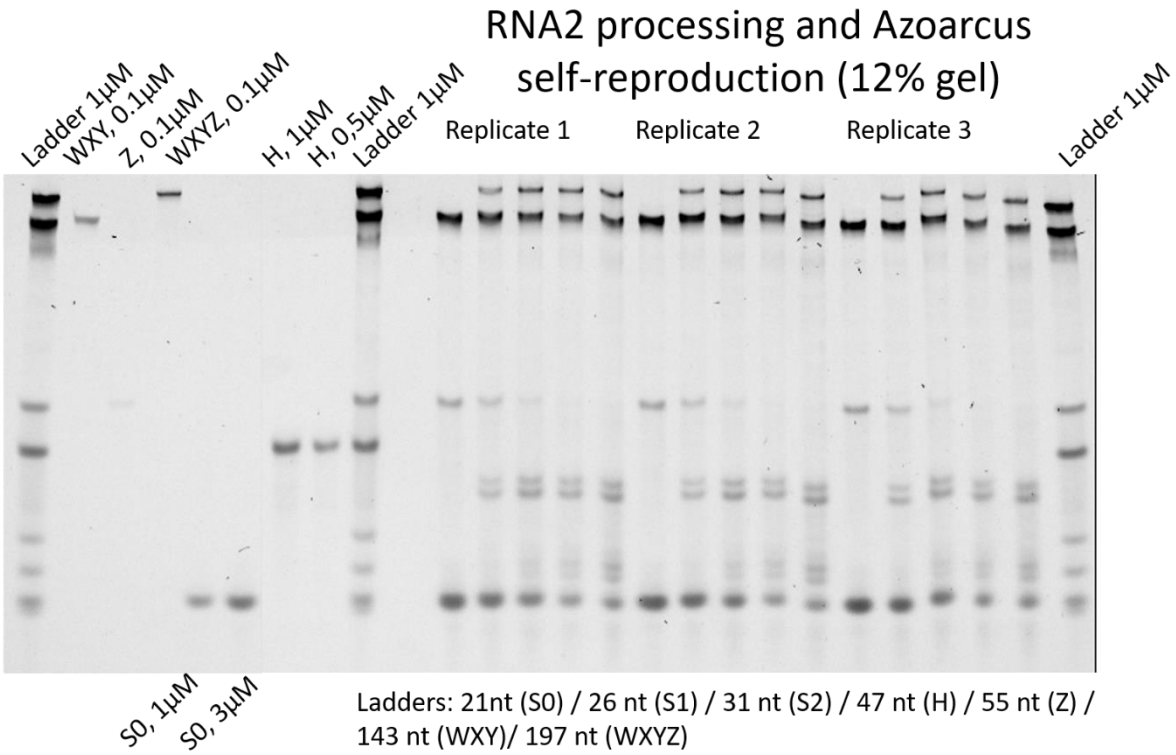

**Gel 5.** RNA3 (5 µM) processing and *Azoarcus* ribozyme self-assembly from WXY (0.5 µM) and Z (0.5 µM) RNA fragments (gel ran to visualize RNA3's products). The time points for Replicate 1, Replicate 2 and Replicate 3 are from left to right: 0h, 1h, 3h, 5h, 7h.




**Gel 6.** 12% acrylamide gel of RNA3 (5 µM) processing and *Azoarcus* ribozyme self-assembly and WXY (0.5 µM) with Z (0.5 µM) RNA fragments (gel ran to visualize WXYZ self-assembly). The time points for Replicate 1, Replicate 2 and Replicate 3 are from left to right: 0h, 1h, 3h, 5h, 7h.




**Gel 7.** 12% acrylamide gel of *Azoarcus* ribozyme self-assembly and WXY (0.5 µM) with Z (0.5 µM) RNA fragments. The time points for Replicate 1, Replicate 2 and Replicate 3 are from left to right: 0h, 1h, 3h, 5h, 7h.



**Gel 8.** 12% acrylamide gel of RNA2 (5 µM) processing and *Azoarcus* ribozyme self-assembly and WXY (0.5 µM) with Z (0.5 µM) RNA fragments. The time points for Replicate 1, Replicate 2 and Replicate 3 are from left to right: 0h, 1h, 3h, 5h, 7h.

# Full details of mathematical models

**Model for RT (RNA1 processing)**

The RT reaction is modelled by the equation: $S_n + S_m \rightarrow S_{n+p} + S_{m-p}$, with rate $k$, which is assumed to be independent of *n*, *m* and *p*. $n$ and $m$ are the numbers of mobile units carried by a stem S, and $p$ is the number of transferred mobile units. In addition, we implement the potential loss of mobile units by hydrolysis when linked to the ribozyme. This effective removal is represented by the reaction: $S_n \rightarrow S_{n-p}$, assumed to occur at the rate $\delta$ irrespective of the number of mobile units $p$. Finally, the slight ribozyme-catalyzed decay of the stems is considered and occurs at the rate $\delta'$: $S_n \rightarrow \emptyset$. Although the kinetic rates of these reactions, $k$, $\delta$ and $\delta'$, depend on the amount of the ribozyme, we assume the rates are constant since the ribozyme concentration does not change significantly in the time scale of the experiments. The rate equation for the dynamics of $S_n$ concentration, $s_n$ is

$$\dot{s}_n = k \sum_{\substack{i+j \geq n \\ i \neq n}} s_i s_j - k s_n (n\, s + \gamma) - \delta \left( n\, s_n - \sum_{i=n+1}^{\infty} s_i \right) - \delta' s_n,$$

where the total mass of stems is $s = \sum_{i=0}^{\infty} s_i$ and of mobile units is $\gamma = \sum_{i=1}^{\infty} i\, s_i$. The first and second terms represent the production and consumption rate of $s_n$ by the transfer reactions, the third term represents the degradation of mobile units, and the last term represents the decay of the stem. We adjusted the parameters $k$ and $\delta$ to fit the species concentrations with the gel electrophoresis data for each time point. We determined the $\delta'$ parameter from a separate experiment: the degradation of **RNA1**'s stem ($S_0$; 5 µM) in the presence of *Azoarcus* ribozyme (0.5 µM) in the same conditions (44°C, 60 mM $M_gCl_2$) as the other experiments of the study. Even though the kinetic model has only 3 parameters, the time course of the amount of the species fits the gel data well (**Fig. 2b**, the fitted parameters are as follows: $k$ = 0.11±0.01 $\mu M^{-1}h^{-1}$, $\delta$ = 0.39±0.03 $h^{-1}$ and $\delta'$ = 0.093±0.007 $h^{-1}$). In the time course, first $s_0$ and $s_2$ increase as $s_1$ decreases, then subsequently all species except $s_0$ decrease to eventual extinction, and finally $s_0$ also gradually decreases.

Here, we briefly discuss the trends in time course, and the exponential distribution in the sequence frequency. Under the assumption that only $S_1$ exists at the initial condition, and using the expression:

$$s_n(t) = s_1(t)(1 - a(t))^{n-1} \; (n \geq 1),$$

we can reduce the rate equations into that of only three variables $s_0(t)$, $s_1(t)$ and $a(t)$,

$$\dot{s_0} = k\frac{s_1}{a^2}(s_1 - (1-a)s_0) + \delta\frac{s_1}{a} - \delta' s_0,$$

$$\dot{s_1} = -s_1\left(k\frac{2}{a}(s_1 - (1-a)s_0) + \delta\left(2 - \frac{1}{a}\right) + \delta'\right),$$

$$\dot{a} = -k(s_1 - (1-a)s_0) + \delta(1-a),$$

where the initial conditions are $a(0) = 1$ and $s_0(0) = 0$.

These rate equations have three parameters $k$, $\delta$ and $\delta'$, which are associated with three trends in the dynamics. The terms with coefficient $k$ due to the transfer reaction in the equations explain the initial increase of $s_0$ and the decrease of $s_1$ and $a$ (i.e., the increase of $s_2 = s_1(1-a)$), until they reach close to the partial equilibrium $s_1 \sim (1-a)s_0$. Next, after the terms with coefficient $\delta$ become dominant, due to degradation of mobile units, $a$ increases to 1 (i.e., $s_2$ decreases to extinction). Finally, when the terms with $\delta'$ become dominant due to the decay of stems, then $s_0$ starts to decrease. Thus, to fit the first trend the model requires only the reaction which transfers mobile units, while to fit the latter two trends it requires two types of degradation reactions, respectively.

Also, the time dependent solution for the equations, $s_n(t)$ $(n \geq 1)$, explains the sequence frequency ($S_0$ ~ $S_7$) showing the exponential distribution in **Fig. 2c**. And its decay factor (at t=45 min), i.e., $\ln(1 - a(t))$ can be predicted from the kinetic parameters ($k, \delta$ and $\delta'$) which were obtained from only the time course data of $S_0$, $S_1$ and $S_2$ (**Fig. 2b**); the decay factor in the sequencing data ($S_1$ ~ $S_7$) is measured as -1.569$\pm$0.057, which well agrees with the model prediction: $\ln(1 - a(t = 45\ \text{min}))$ = -1.589.

**Model for TSA (RNA2 processing)**

Production of hairpin and circular RNAs by the TSA mechanism are depicted by the following reactions: $S + S \rightarrow H$, and $H \rightarrow C$, at the rates $k_1$ and $k_2$. We assume the degradation of the RNA species is negligible in the time scale of experiments due to the presence of more stable hairpins and cyclic RNA species. The rate equations for their concentrations, $s, h$ and $c$ are:

$$\dot{s} = -k_1 s^2,\ \dot{h} = \frac{1}{2}k_1 s^2 - k_2 h,\ \text{and}\ \dot{c} = k_2 h,$$

respectively. Assuming the initial condition with only S present, these differential equations can be solved explicitly as functions of time:

$$s(t) = \frac{s(0)}{1+k_1\, s(0)\, t},\ h(t) = \frac{s(0)}{2}\left(A(t) - \frac{1}{1+k_1\, s(0)\, t}\right),\ \text{and}\ c(t) = \frac{s(0)}{2}(1 - A(t)),$$

where $A(t) = e^{-k_2 t}\left(1 + \int_0^t \frac{k_2 e^{k_2 t'}}{1+k_1 s(0)\, t'} dt'\right) \sim \frac{1}{1+k_1\, s(0)t} + e^{-k_2 t}\frac{k_1 s(0) t}{1+k_1 s(0) t}$. Note that the total mass of stems is conserved: $s(t) + 2h(t) + 2c(t) = s(0)$. As before, we also adjusted the parameters to fit the concentrations of S, H and C predicted by these rate equations to the gel electrophoresis data. For the constraint on the conservation of stems, we introduce an additional calibration multiplying the values for H and C by the same constant $\xi$, which is also fitted along with the other parameters. $\xi$ corrects the fact that H and C gel bands are wider than single-sequence bands, since there are two attack sites for TSA (after the second and fourth G according to sequencing data). The fitted parameters values (**Fig. 3b**) are as follows: $k_1$ = 0.20±0.02 $\mu$M$^{-1}$h$^{-1}$, $k_2$ = 0.24±0.01 h$^{-1}$ and $\xi$=0.59±0.01. The solutions $s(t)$, $h(t)$ and $c(t)$ well explain the time course of the gel data: the first rapid decrease of $s(t)$ from the initial $s(0)$, the increase followed by the decrease of $h(t)$, and the final convergence of $c(t)$ to $\frac{s(0)}{2}$.

**Model for RT + TSA (RNA3 processing)**

Here, as before, we assume the RT mechanism, which transfers a mobile unit between molecules, works at the same rate $k$ for any $S_n$ and irrespective of the number of mobile units at the 3'-end of a hairpin RNA $H_{m,n}$. We also assume the degradation of a mobile unit during the recombination reaction occurs at the rate $\delta$, and the catalyzed decay of a stem occurs at the rate $\delta'$. Similarly, the TSA mechanism works for any $S_m$ and $S_n$ at the rate $k_1$ or any $H_{m,n}$ at the rate $k_2$. The rate equations for the concentrations of $S_n$ and $H_{m,n}$ noted as $s_n$ and $h_{m,n}$ are formally:

$$\dot{s}_n = k\sum_{\substack{i+j\geq n \\ i\neq n}} s_i\left(s_j + h_j\right) - k s_n(n(s+h) + \gamma + \gamma_h) - \delta\left(n\, s_n - \sum_{i=n+1}^{\infty} s_i\right) - \delta' s_n - k_1 s_n s,$$

$$\dot{h}_{m,n} = k\sum_{\substack{i+j\geq n \\ i\neq n}} h_{m,i}\left(s_j + h_j\right) - k h_{m,n}(n(s+h) + \gamma + \gamma_h) - \delta\left(n\, h_{m,n} - \sum_{i=n+1}^{\infty} h_{m,i}\right) - 2\delta' h_{m,n} + \frac{1}{2}k_1 s_m s_n - k_2 h_{m,n},$$

where $h_n$ is the total of hairpin RNA with $n$ available mobile units: $h_n = \sum_{i=0}^{\infty} h_{i,n}$, $h$ is the total of the all hairpin RNAs: $h = \sum_{i=0}^{\infty} h_i$, and $\gamma_h$ is the total mass of available mobile units within all the hairpins: $\gamma_h = \sum_{i=1}^{\infty} i\, h_i$. In both of the equations, the first terms are analogous to the ones used in the previous RT model, and the last terms are the ones used for the previous TSA model.

In this kinetic model, there are the five parameters: $k$, $k_1, k_2$, $\delta$ and $\delta'$ coming from the fusion of the two previous models. The fitted parameter values (**Fig. S5**) are as follows: $k$ = 0.084±0.009 $\mu$M$^{-1}$h$^{-1}$, and $k_1$ = 0.018±0.002 $\mu$M$^{-1}$h$^{-1}$, $k_2$ = 1.0±0.2 h$^{-1}$, $\delta$ = 0.21±0.03 h$^{-1}$ and $\delta'$ = 0.13±0.01 h$^{-1}$. The WXY+Z+**RNA3** theoretical plot (**Fig. S8**) is obtained by the same fitting procedure as for RNA3 only.

The fitted parameter values (**Fig. S8**) are: $k$ = 0.06±0.01 $\mu$M$^{-1}$h$^{-1}$, and $k_1$ = 0.010±0.001 $\mu$M$^{-1}$h$^{-1}$, $k_2$ = 0.62±0.16 h$^{-1}$, $\delta$ = 0.17±0.03 h$^{-1}$ and $\delta'$ = 0.11±0.01 h$^{-1}$. All of the parameters are reduced by 20~60% due to the reduction of the ribozyme activity.

Here, we formally discuss that the time dependent solution $s_n(t)$ and $h_{m,n}(t)$ are indeed the combination of the solutions for the rate equations in the previous models (for RT and TSA): First, note that $s(t)$ and $h(t)$ obey the rate equations $\dot{s} = -k_1 s^2 - \delta' s$ and $\dot{\mathrm{h}} = \frac{1}{2}k_1 s^2 - (k_2 + 2\delta')h$, which are the same as that for TSA except the decay term, and, in principle, can be solved explicitly.
Second, applying the same procedure as we did for the model for RT, using the expression $\hat{s}_n = s_n + h_n$, and $\hat{s}_n(t) = \hat{s}_1(t)(1-\hat{a}(t))^{n-1}$ $(n \geq 1)$, then, the rate equations for $s_n$ and $h_n$ are reduced to:

$$\dot{\hat{s}_0} = k\frac{\hat{s}_1}{a^2}(\hat{s}_1 - (1-\hat{a})\hat{s}_0) + \delta\frac{\hat{s}_1}{\hat{a}} - D\hat{s}_0,$$

$$\dot{\hat{s}_1} = -\hat{s}_1\left(k\frac{2}{a}(\hat{s}_1 - (1-\hat{a})\hat{s}_0) + \delta\left(2 - \frac{1}{\hat{a}}\right) + D\right),$$

$$\dot{\hat{a}} = -k(\hat{s}_1 - (1-\hat{a})\hat{s}_0) + \delta(1-\hat{a}),$$

where $D = (\delta'(2-\alpha) + \frac{1}{2}k_1\alpha^2(\hat{s}_0 + \frac{\hat{s}_1}{\hat{a}}) + k_2(1-\alpha))$, and we define $\alpha(t) = \frac{s(t)}{s(t)+h(t)}$, and then $s_n(t) = \alpha(t)\hat{s}_n(t)$ and $h_n(t) = (1-\alpha(t))\,\hat{s}_n(t)$. That is, we obtain the same rate equations as the model for RT except for the terms with $\alpha$, which depends on $t$.

Lastly, we decompose $h_{m,n}(t)$ as $h_{m,n}(t) = \beta_m(t)h_n(t)$, where $\beta_m(t)$ obeys the differential equation:

$$\dot{\beta_m} = \frac{k_1 s}{2h}(s_m - \beta_m s).$$

Therefore, the time dependent solutions $s_n(t)$ and $h_{m,n}(t)$ are constructed as:

$$s_n(t) = \alpha(t)\big(1-\hat{a}(t)\big)^{n-1}\hat{s}_1(t) \qquad (n \geq 1),$$

$$h_{m,n}(t) = \beta_m(t)\big(1-\alpha(t)\big)\big(1-\hat{a}(t)\big)^{n-1}\hat{s}_1(t)\ (n \geq 1).$$

Therefore, $h_{m,n}$ and $s_n$ exponentially decrease with $n$ with the same decay factor, $\ln\big(1-\hat{a}(t)\big)$. $\beta_m$ and therefore $h_{m,n}$ also decrease with $m$, but not exponentially.

As in the case of **RNA1** processing, the sequence frequencies for $S_n$ and $H_{m,n}$ (at t=45 min; **Fig. 4b**) can be predicted from the five kinetic parameters ($k, \delta, \delta'$, $k_1$ and $k_2$) which were obtained from only the time course data of $S_0$, $S_1$, $S_2$ and H measured on gel (**Fig. S5**); the decay factor for $S_n$ in the sequencing data ($S_1$ ~ $S_7$) is measured as -1.74±0.09, which is close to the model prediction: -1.650. Further, the decay factor for $H_{m,0}$~$H_{m,3}$ for each $m$ is measured as -1.33±0.01, which is in the range of the measured and predicted ones for $S_n$ as is expected by the above analysis of the kinetic model. The

distribution in sequencing data and model predictions are, overall, in good agreement, although there are deviations (up to one order of magnitude) at $H_{m,0}$ and $H_{0,m}$ ($0 \leq m \leq 3$); the empirical values are larger for $H_{m,0}$, while the predicted values are larger for $H_{0,n}$. A possible explanation is a difference in the kinetic rates for hairpin formation at the end of a dimer with and without mobile units. Extremities with mobile units may form hairpins more easily than those without due to steric effects. This would cause $H_{0,m}$ to formed more slowly and experimentally lead to less products than predicted. Similarly, the formation of $C_{m,0}$ from $H_{m,0}$ is also slower, and $H_{m,0}$ would be experimentally larger than predicted.

**Table S2. Parameters summary with covalent ribozymes**

| | $\delta$ ($h^{-1}$) | $\delta'$ ($h^{-1}$) | $k$ ($\mu M^{-1}h^{-1}$) | $k_1$ ($\mu M^{-1}h^{-1}$) | $k_2$ ($h^{-1}$) |
|---|---|---|---|---|---|
| RT only | 0.39 (0.03) | 0.093 (0.007) | 0.11 (0.01) | | |
| TSA only | | | | 0.20 (0.02) | 0.241 (0.009) |
| RT + TSA | 0.21 (0.03) | 0.13 (0.01) | 0.084 (0.009) | 0.018 (0.002) | 1.0 (0.2) |

The values in parenthesis are the standard errors for the parameters (see model fitting procedure)

**Table S3. The fitted parameters for RNA3 processing when seeded with WXY + Z:**

| | $\delta$ ($h^{-1}$) | $\delta'$ ($h^{-1}$) | $k$ ($\mu M^{-1}h^{-1}$) | $k_1$ ($\mu M^{-1}h^{-1}$) | $k_2$ ($h^{-1}$) |
|---|---|---|---|---|---|
| RT + TSA | 0.17 (0.03) | 0.11 (0.01) | 0.06 (0.01) | 0.010 (0.001) | 0.*62* (0.16) |

The RT + TSA experiment is not deducible from the fitted parameters from the RT and TSA only experiments. The predictive value of the models relies on the consistency with the sequencing data.

$k_1$ and $k_2$ correspond to the kinetic rates of the formation of hairpins and cyclic RNAs respectively. In the TSA only experiment (**RNA2**), only $H_{0,0}$ and $C_{0,0}$ species are involved. In the TSA + RT experiment (**RNA3**), $H_{i,j}$ and $C_{i,j}$ (with i and j going from 0 to at least 3) are involved. In the latest case $k_1$ and $k_2$ correspond to global kinetic rates considering different types of TSA reaction that might have different kinetic rates (depending on the numbers of mobile units involved). This explains why $k_1$ and $k_2$ are different in TSA only and in RT + TSA experiments.

**Model for WXYZ self-reproduction, with and without RNA3**

Here, we assume that the complex formation WXY + Z → WXY:Z is very fast[11,13] and all WXY and Z species are included in the non-covalent, yet catalytic, complex WXY:Z. The model takes into account the two reversible reactions WXYZ + WXY:Z → WXYZ + WXYZ and WXY:Z + WXY:Z → WXYZ + WXY:Z with the forward rates $k_a$ and $k_b$, and the backward rates $rk_a$ and $rk_b$, respectively. Here, we assume that the catalytic efficiencies for covalent and non-covalent ones are the same, $k_a = k_b$, as an approximation of observed rates in previous studies.[9,14] We assume the conservation of the total mass of WXY; thus, the concentration of WXY:Z is $a_0 - a$, where $a$ is the concentration of WXYZ, and $a_0$ is the initial concentration of WXY. Therefore, the rate equations for WXYZ is, $\dot{a} = a_0 k_a\big((a_0 - a) - ra\big)$, which has the solution $a(t) = \frac{a_0}{1+r}(1 - e^{-a_0 k_a (1+r)t})$. In the data fitting we fitted the concentrations of WXYZ predicted by the rate equation with the corresponding gel electrophoresis data (**Fig. S9**). The fitted parameters in the absence of RNA3 species are $k_a = 0.97 \pm 0.07$ $\mu M^{-1}h^{-1}$ and $r = 1.0 \pm 0.1$. Because RNA3 has a tag, it competes with Z species to bind to WXY's IGS, thus acting as an inhibitor. The fitted parameters in the presence of RNA3 species are $k_a = 0.25 \pm 0.02$ $\mu M^{-1}h^{-1}$ and $r = 1.3 \pm 0.1$.

**Model fitting procedure**

In fitting each model to the corresponding electrophoresis data, we optimized the parameter set $\boldsymbol{\alpha}$ (e.g., $\boldsymbol{\alpha} = \{k, k_1, k_2, \delta, \delta'\}$ in the model for RT+TSA) to minimize the weighted sum square residue:

$$\chi^2 = \sum_i \sum_j \left( \frac{y_{i,j} - y_i(t_j|\boldsymbol{\alpha})}{\sigma} \right)^2,$$

Where $\sum_i$ is the sum over all species and $\sum_j$ is the sum over all time points $t_j = 1,3,5$ and 7, $y_{i,j}$ is the data value for the species $i$ at time $t_j$, and $y_i(t|\boldsymbol{\alpha})$ is the model prediction for $i$ using $\boldsymbol{\alpha}$ (e.g. $s_0, s_1, s_2$ and $h$ for RT+TSA)**.** We weight each error between a data point and the corresponding model prediction by $\sigma = \max_i(\sigma_i)$, the maximum of the standard error obtained from the triplicate among all points for the species $i$, in order to avoid giving extra importance to data points which had a low standard deviation. The standard error for the fitted parameter $\alpha_i$ is computed as $\pm\sqrt{C_{ii}}$, where the covariance matrix $C$ is the inverse of the half Hessian matrix H, $H_{ij} = \frac{1}{2}\partial_{\alpha_i}\partial_{\alpha_j}\chi^2$.[15]

# Bibliography


1 W. E. Draper, E. J. Hayden and N. Lehman, *Nucleic Acids Res.*, 2008, **36**, 520–531.

2 A. J. Zaug, M. M. McEvoy and T. R. Cech, *Biochemistry*, 1993, **32**, 7946–7953.

3 A. J. Zaug and T. R. Cech, *Science*, 1985, **229**, 1060–1064.

4 M. D. Been and T. R. Cech, *Science*, 1988, **239**, 1412–1416.

5 B. M. Chowrira, A. Berzal-Herranz and J. M. Burke, *EMBO J.*, 1993, **12**, 3599–3605.

6 Q. Vicens and T. R. Cech, *Nat. Chem. Biol.*, 2009, **5**, 97–99.

7 H. F. Tabak, G. Van der Horst, A. M. J. E. Kamps and A. C. Arnberg, *Cell*, 1987, **48**, 101–110.

8 J. L. Litke and S. R. Jaffrey, *Nat. Biotechnol.*, 2019, **37**, 667–675.

9 E. J. Hayden and N. Lehman, *Chem. Biol.*, 2006, **13**, 909–918.

10 S. Arsène, S. Ameta, N. Lehman, A. D. Griffiths and P. Nghe, *Nucleic Acids Res.*, 2018, **46**, 9660–9666.

11 S. Ameta, S. Arsène, S. Foulon, B. Saudemont, B. E. Clifton, A. D. Griffiths and P. Nghe, *Nat. Commun.*, 2021, **12**, 842.

12 J. J. Goetz and J. M. Trimarchi, *Nat. Biotechnol.*, 2012, **30**, 763–765.

13 N. Vaidya, M. L. Manapat, I. A. Chen, R. Xulvi-Brunet, E. J. Hayden and N. Lehman, *Nature*, 2012, **491**, 72–77.

14 E. J. Hayden, G. Von Kiedrowski and N. Lehman, *Angew. Chemie - Int. Ed.*, 2008, **47**, 8424–8428.

15 W. H. Press, S. A. Teukolsky, W. T. Vetterling and B. P. Flannery, *Numerical Recipes 3rd Edition: The Art of Scientific Computing*, Cambridge Universiity Press, Cambridge, 2007.