μIVC-Useq: a microfluidic-assisted high-throughput functionnal screening in tandem with next generation sequencing and artificial neural network to rapidly characterize RNA molecules

Roger Cubi, Farah Bouhedda, Mayeul Collot, Andrey Klymchenko, Michaël Ryckelynck

METHOD

μIVC-Useq: a microfluidic-assisted high-throughput functional screening in tandem with next-generation sequencing and artificial neural network to rapidly characterize RNA molecules

ROGER CUBI,[1,3] FARAH BOUHEDDA,[1,3] MAYEUL COLLOT,[2] ANDREY S. KLYMCHENKO,[2] and MICHAEL RYCKELYNCK[1]

[1]Université de Strasbourg, CNRS, Architecture et Réactivité de l'ARN, UPR 9002, F-67000 Strasbourg, France
[2]Université de Strasbourg, CNRS, Laboratoire de Bioimagerie et Pathologies, UMR 7021, 67401 Illkirch, France

ABSTRACT

The function of an RNA is intimately linked to its structure. Many approaches encompassing X-ray crystallography, NMR, structural probing, or in silico predictions have been developed to establish structural models, sometimes with a precision down to atomic resolution. Yet these models still require experimental validation through the preparation and functional assay of mutants, which can rapidly become time consuming and laborious. Such limitations can be overcome using high-throughput functional screenings that may not only help in validating the model, but also inform on the mutational robustness of a structural element and the extent to which a sequence can be modified without altering RNA function, an important set of information to assist RNA engineering. We introduced the microfluidic-assisted in vitro compartmentalization (μIVC), an efficient and cost-effective screening strategy in which reactions are performed in picoliter droplets at rates of several thousand per second. We later improved μIVC efficiency by using it in tandem with high-throughput sequencing, though a laborious bioinformatic step was still required at the end of the process. In the present work, we further increased the automation level of the pipeline by implementing an artificial neural network enabling unsupervised bioinformatic analysis. We demonstrate the efficiency of this "μIVC-Useq" technology by rapidly identifying a set of sequences readily accepted by a key domain of the light-up RNA aptamer SRB-2. This work not only shed some new light on the way this aptamer can be engineered, but it also allowed us to easily identify new variants with an up to 10-fold improved performance.

Keywords: droplet-based microfluidics; high-throughput screening; bioinformatics; RNA engineering; light-up RNA aptamer

INTRODUCTION

RNA is able to perform a wide range of functions (e.g., scaffolding, recognition, or catalysis) that are intimately linked to the three-dimensional architecture of the molecule and its capacity to properly display key residues in space. Whereas, in general, interaction with macromolecules (e.g., proteins or nucleic acids) does not necessarily involve high structural complexity, on the contrary, the tight and specific recognition of small ligands usually occurs through sophisticated structures best exemplified by those found in riboswitches aptamer domains (Winkler and Breaker 2005; Pavlova et al. 2019; Sherlock and Breaker 2020; Husser et al. 2021). The fine understanding of the recognition mechanism at work is greatly facilitated by the knowledge of the tridimensional structure of the RNA complexed to its ligand using NMR or X-ray crystallography (Ennifar 2016; Turner and Mathews 2016). Besides, faster, yet less precise, structural data can be obtained from probing in solution (Turner and Mathews 2016) or in silico folding prediction (Zuker 2003; Zadeh et al. 2011). In any case, a structural model in which RNA adopts a folding made of the combination of a variety of structural elements (e.g., stems, loops or more complex tertiary

[3]These authors contributed equally to this work.
Corresponding author: m.ryckelynck@unistra.fr
Article is online at http://www.rnajournal.org/cgi/doi/10.1261/rna.077586.120.

interactions) is generated. Yet, it usually requires an experimental validation through the preparation and functional evaluation of mutants. Advantageously, such mutagenic analysis may help in confirming the key role of some residues establishing specific contacts and that are expected to have a low tolerance to mutations. Moreover, it may also give access to data like, for instance, the mutational robustness of the different elements (i.e., the capacity of an element to tolerate mutations and/or sequence permutations). Indeed, even for an element as simple as a stem, it is difficult to precisely predict what impact (beneficial, negative, or neutral) the conservative mutation of base pairs may have on RNA functionality. Nonetheless, such information would be extremely valuable to assess the mutational robustness of a molecule or even to assist its engineering.

An efficient way to evaluate mutation tolerance and engineerability of an RNA domain consists in looking at nucleotide conservation through the alignment of its orthologs. Then, using such alignment, one can evaluate the tolerance of each position to variability by scoring the conservation of each residue (e.g., using Shannon uncertainty [Schneider et al. 1986]) and even compute the informational complexity of the molecule (Carothers et al. 2004). As a matter of fact, the more sequences are used to generate the alignment, the more accurate the predictions will be. The situation may slightly differ with the origin of the RNA. On the one hand, when studying natural RNAs, large sets of sequences can rapidly be collected by browsing the vast reservoir of genomes. A detailed view of a sequence evolution can then be obtained by its phylogenetic study. Such comparative genomic analysis is typically used to identify and characterize riboswitches (Barrick et al. 2004; Weinberg et al. 2010, 2017). As a first approximation, one may consider that finding a sequence evolutionarily conserved in a genome indicates that the molecule is likely functional. Then, searching for covariations enables to rapidly highlight putative secondary structure elements, while looking at nucleotide conservation allows us to anticipate which residues are important for the function of the molecule (e.g., ligand recognition) as they are expected to be highly conserved. On the other hand, artificial RNAs (e.g., RNA aptamers or ribozymes) can rapidly be isolated using in vitro selection procedures like SELEX (Ellington and Szostak 1990; Tuerk and Gold 1990), at the end of which several sequence families are typically identified. The best prototype sequence can then possibly be refined through a so-called doped SELEX step, during which the sequence is partially randomized prior to being subjected to a few rounds of SELEX to isolate the best fit sequences (Bartel et al. 1991). This strategy is particularly efficient to characterize binders (i.e., aptamers), but it is more questionable for RNAs endowed with more complex functions like catalysis or fluorescence activation since a capacity to bind a target is not necessarily synonymous with efficient substrate conversion or fluorescence emission. Instead, such RNAs are expected to be more efficiently developed by using SELEX in tandem with functional screening, where each sequence contained in a library is individually assayed for the target function and sorted from the bulk accordingly (Autour and Ryckelynck 2017). This way, large libraries can be rapidly enriched in variants of interest using the capacity of SELEX to handle as many as 10^{15} different molecules at once, prior to using the accuracy of functional screening to select the best adapted variants (Filonov et al. 2014; Bouhedda et al. 2020).

To be viable and competitive, a screening technology should be rapid, cost-effective and operate in a high-throughput manner while offering the best possible control over reaction conditions. In this view, microfluidics is particularly attractive as it both allows us to significantly decrease reaction volumes while increasing analytical throughputs. For instance, several hundreds of sequences can be individually expressed and analyzed in parallel using a microfluidic chip in which several hundreds of nanoliter volume micro-compartments are fabricated (Ketterer et al. 2015). Such large-scale integration devices are interesting to exhaustively screen libraries made of a few hundred different sequences. Substantial gain in the throughput can be achieved by repurposing high-throughput sequencing machines and using their sequencing microfluidic flow-cells to sequence immobilize variant coding genes, sequence them prior to expressing them and measuring their phenotype (Denny and Greenleaf 2019). Yet, this requires a high-throughput sequencing platform to be customized and it is not easily applicable to devices currently available. An alternative affording similar, if not higher, analysis throughput exploits emulsion-based technologies like particle display (Wang et al. 2014) or droplet-based microfluidics (Autour and Ryckelynck 2017). These technologies allow the parallel analysis of several million variants in a quantitative way. Over the past years, we demonstrated that microfluidic-assisted in vitro compartmentalization (μIVC) is extremely efficient at identifying optimized ribozymes (Ryckelynck et al. 2015) and light-up RNA aptamers (Autour et al. 2016, 2018; Bouhedda et al. 2020). Moreover, monitoring the fate of individual sequences throughout the screening using next-generation sequencing (NGS) allowed us to identify biosensors with optimized communication modules, though a relatively basic and time-consuming sequence analysis was performed (Autour et al. 2019). In this paper, we introduce a new technological advance called μIVC-Useq, in which μIVC is used in tandem with NGS and an artificial neural network (ANN) algorithm to rapidly profile large libraries and identify all sequence permutations tolerated by an RNA domain in an unsupervised manner. Using this technology, we were able to exhaustively evaluate the performances of ~1,000,000 sequence permutations of a putative stem contained in an aptamer, prior to reducing

the information to only two clusters of sequences from which we were able to confirm the predicted structure as well as to identify ~10-fold improved mutants that would have been impossible to identify a priori by simply using structural criteria. We believe that the efficiency of μIVC-Useq together with its high degree of automation represents a major advance in the field by both allowing us to evaluate the robustness of RNA domains and assisting their engineering.

RESULTS

SRB-2 aptamer as a model aptamer

To set up and evaluate the potential of our technology to identify variants of interest in an unsupervised manner, we chose SRB-2 aptamer (Fig. 1A) as a model RNA (Holeman et al. 1998). This aptamer was originally isolated for its capacity to interact with the sulforhodamine B dye. Later, SRB-2 was also found to be able to activate fluorogenic forms of the sulforhodamine B (Sunbul and Jaschke 2013) as well as close homologs (Sunbul and Jaschke 2018) made of a dye conjugated to a dinitroaniline moiety inducing a contact quenching. Moreover, we recently expanded the set of SRB-derived fluorogenic dyes by developing Gemini-561 (Fig. 1B,C), a dimerized form of sulforhodamine B that self-quenches by forming H-aggregates (Bouhedda et al. 2020). While Gemini-561 is optimally activated by the light-up aptamer o-Coral (an SRB-2 derived aptamer), it can also form a fluorescent complex with SRB-2 aptamer to some extent.

From a structural point of view, the minimal form of SRB-2 was proposed to fold into three stems (P1, P2, and P3) closed by apical loops L2 and L3 and spaced by a long unpaired stretch J2–3 (Fig. 1A). Previous studies showed that P2 and L2 can be modified without compromising the function of the aptamer (Holeman et al. 1998). Moreover, we recently showed that the sequence of P1 can be readily modified, provided a stem can still form (Bouhedda et al. 2020). The remaining elements J2/3, P3, and L3 were originally predicted to encompass the sulforhodamine B-binding site (Holeman et al. 1998). It is quite likely that J2/3 and L3 adopt a rather elaborated tridimensional structure stabilized by tertiary interactions, that poorly tolerates mutations. Properly deciphering the intimate organization of this region of the dye-binding will require a long and dedicated structural characterization using X-ray crystallography or NMR characterization. Finally, whereas the existence of a P3 stem has been supported by the identification of sequence covariation (Holeman et al. 1998), the actual tolerance of this region to sequence permutation as well as its degree of optimization have never been studied, making it an interesting model to validate our technology. Therefore, to shed further light on this region, we prepared a mutant library in which 10 out of the 12 nt of this putative P3 stem were randomized (Fig. 1A). We choose to limit our analysis to 10 positions to be able to screen the resulting library as exhaustively as possible (the μIVC screening being currently limited to the analysis of 10 million mutants per experiment). We reasoned that the rather simple organization of this region should ease interpretations and validation of our approach while allowing us to determine the minimal length P3 should adopt, identify eventual sequence biases and, perhaps, sequences displaying better properties than the wild-type parental molecule.

Functional screening of the libraries

The P3 mutant library was obtained by chemical synthesis. The region encoding SRB-2-derived aptamer was

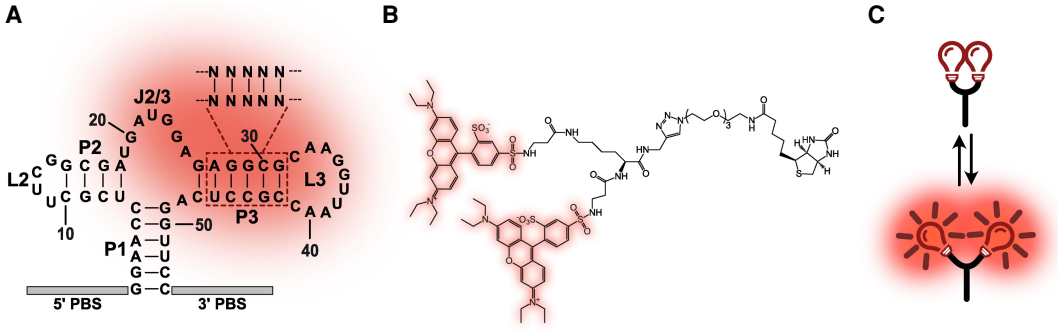


FIGURE 1. Structure and mode of action of SRB-2 RNA aptamer and Gemini-561 fluorogen. (*A*) SRB-2 aptamer and P3 mutant library. The molecule is shown as initially described (Holeman et al. 1998). The 10 positions randomized in the P3 stem are dashed-boxed and the domains proposed to interact with sulforhodamine B are shadowed in red. SRB-2 and its derivatives were flanked with two constant regions used as primer binding sequences (PBS) for PCR amplifications. (*B*) Gemini-561 fluorogen. The fluorogen is made of a dimer of sulforhodamine B (shadowed in red) linked together by a lysine and a PEG linker. The fluorogen also carries a biotin moiety initially used for aptamer selection (Bouhedda et al. 2020). (*C*) Working principle of Gemini-561. H-aggregates form when the molecule is dissolved in an aqueous environment, leading to sulforhodamine B fluorescence quenching. However, modulating medium polarity (e.g., by adding methanol) or in the presence of SRB-2 aptamers, both sulforhodamine B moieties separate and recover their fluorescence capacity.

surrounded by constant regions (used for PCR amplification) and placed downstream from the T7 RNA polymerase promoter. Moreover, a unique droplet identifier (UDI) was appended upstream of the T7 RNA polymerase promoter. By analogy with the unique molecular identifiers (UMIs) used to precisely establish transcript copy-number in transcriptomic analyses (Islam et al. 2014), UDIs are sequences of 20 randomized nucleotides expected to be unique to each molecule of the library (10^{12} different possible sequence permutations with only 10^7 molecules tested at most during a conventional μIVC screening) and later used to cluster together molecules originating from the same droplet (see below). DNA molecules were then diluted into a PCR amplification mixture prior to being individualized into 2.5 pL water-in-oil (W/O) droplets (step i on Fig. 2). Adjusting DNA dilution allowed us to modulate the average starting number of genes per droplet, a value also known as λ, and doing so to control the rate of multiple encapsulation events. Indeed, as DNA molecules distribute into the droplets according to Poisson statistics (Mazutis et al. 2009), knowing λ makes it possible to precisely calculate the fraction of droplets initially occupied by one or more genes (Supplemental Table S1). λ was initially kept high (i.e., λ = 2) to maximize the number of analyzed molecules. Yet, once theoretical sequence diversity decreased, λ was reduced in order to limit multiple encapsulation of several templates and gain in analysis accuracy. W/O droplets were then produced by infusing the aqueous phase into a microfluidic chip together with a fluorinated oil phase supplemented in a fluorosurfactant (Holtze et al. 2008) to stabilize droplets. Upon production, droplets were thermocycled to PCR-amplify each gene into ~300,000 copies.

Then, droplets were reinjected into a droplet fusion microfluidic device (Mazutis et al. 2009) in which they were spaced by a stream of fluorinated oil (supplemented with 2% fluorosurfactant) and synchronized one-to-one with larger (16 pL) droplets generated on-chip and containing an in vitro transcription (IVT) mixture supplemented in Gemini-561 (Bouhedda et al. 2020). Pairs of droplets were then fused together when passing between a pair of electrodes energized by an AC field (step ii on Fig. 2). One-to-one pairing was maximized by monitoring the blue fluorescence (coumarin acetate added into both sets of droplets). Indeed, whereas PCR droplets were highly fluorescent (20 μM of coumarin acetate), IVT ones displayed a much lower signal (1 μM of coumarin acetate). As a consequence, the blue fluorescence of the merged droplets allows us to discriminate unfused IVT droplets (low blue fluorescence; 20–30 RFUs on Fig. 3A,B and 10–20 RFUs on Fig. 3C) from those fused with one (moderate blue fluorescence; 50–70 RFUs on Fig. 3A,B and 25–30 RFUs on Fig. 3C) or even two PCR droplets (high blue fluorescence; 80–90 RFUs on Fig. 3A, 90–100 RFUs on Fig. 3B and 35–45 RFUs on Fig. 3C). In general, more than 86% of

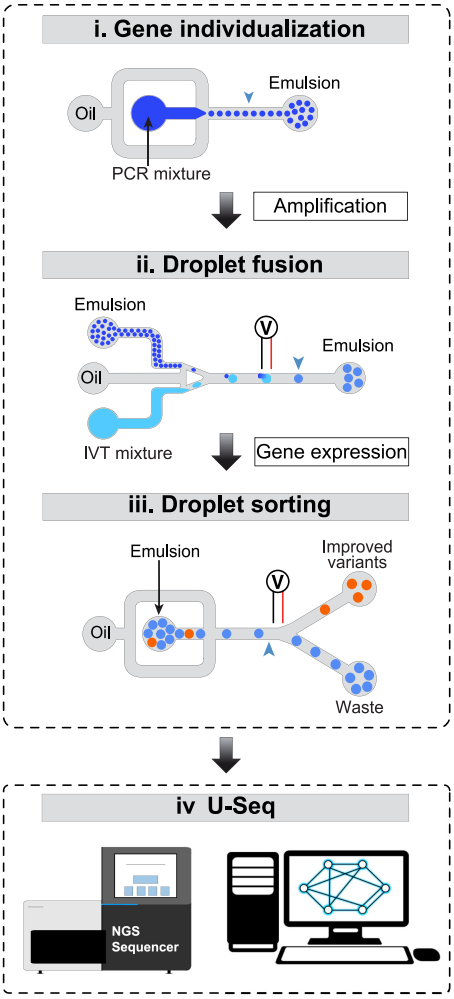


FIGURE 2. Overview of the μIVC-Useq pipeline. The microfluidic-assisted in vitro compartmentalization (μIVC) step is made of three main steps during which: (i) The genes contained in a library are individualized prior to being amplified, (ii) droplets containing amplified genes are fused one-to-one with an in vitro expression mixture supplemented in fluorogen, and (iii) the fluorescence of each droplet is measured and used to sort them accordingly. In μIVC-Useq, the screening step is followed by a next-generation sequencing (NGS) analysis of the sequences contained in enriched libraries while an unsupervised bioinformatic pipeline allows the rapid identification of molecules of interest.

IVT droplets were fused with one PCR droplet (Supplemental Table S1). Upon fusion, droplets were collected and incubated for 2 h at 37°C to allow the genes to be transcribed and, if functional, RNA to complex and activate the orange fluorescence of Gemini-561.

Finally, droplets were reinjected into a fluorescence-activated droplet sorting device (Baret et al. 2009), in which they were spaced by a surfactant-free oil stream prior to getting their fluorescence analyzed (step iii on Fig. 2). Those droplets displaying a blue fluorescence corresponding to single-fused droplets as well as a significant increase

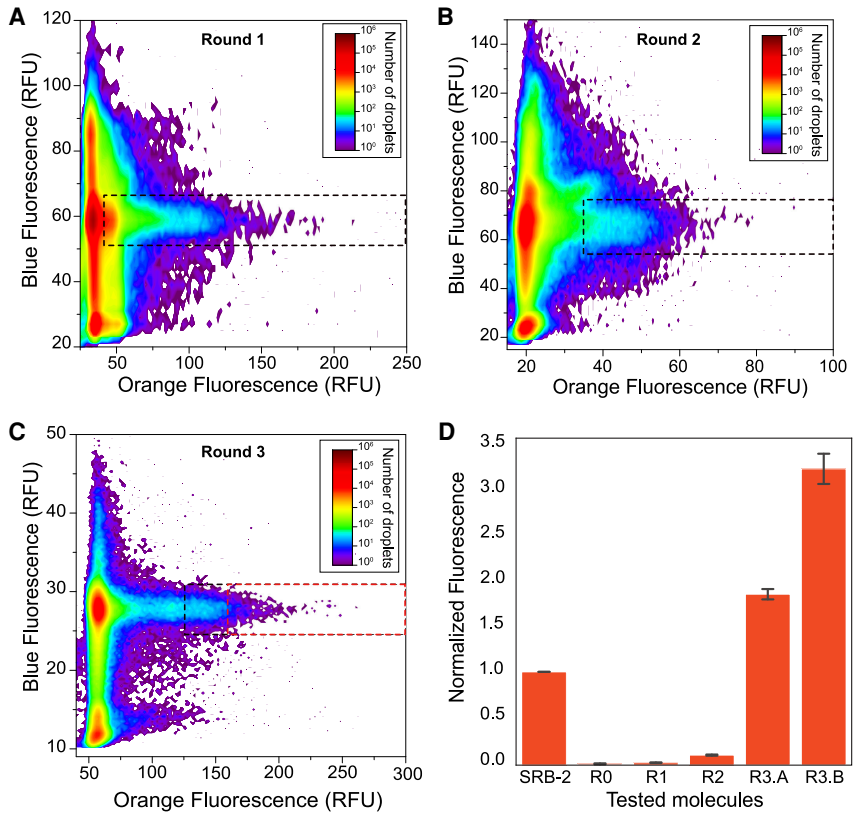


FIGURE 3. Fluorescence profiles of droplets obtained at each round of μIVC. Fusion efficiency can be assessed by the blue fluorescence of coumarin added in the droplets while the orange fluorescence informs on RNA function (Gemini-561 fluorescence activation). The population of droplets gated and sorted during the first (*A*) and the second (*B*) round of screening are dashed-boxed in black. Their DNA content was recovered and used to prime a new round of screening. During the third round (*C*), a relaxed (black dashed box, population R3A) and a stringent (red dashed box, population R3B) gating were used. (*D*) For each round of μIVC, the enriched DNA libraries were in vitro transcribed in the presence of 500 nM Gemini-561 and the fluorescence was monitored over time. The fluorescence apparition rate was computed for each library and normalized to that of the parental SRB-2 aptamer. The values are the mean of $n = 3$ independent experiments. The error bars correspond to ± 1 SD.

of their orange fluorescence (Gemini-561/RNA complex formation) were gated as positive (boxed populations on Fig. 3A–C) and specifically sorted from the bulk. Upon sorting, droplets were broken, their DNA content recovered by PCR and their original UDI was exchanged for a new one to be able to track individual droplets during the next round of screening.

We performed three rounds of such μIVC screening while gently increasing the selection stringency. Indeed, the first round was performed at a high droplet occupancy (Supplemental Table S1) to maximize the number of analyzed molecules, and we selected any droplet displaying an orange fluorescence above the background (Fig. 3A) to limit the loss of molecules of interest. The second round was performed at lower droplet occupancy and only droplets displaying a significant fluorescence were recovered (Fig. 3B). Proper enrichment of the libraries was confirmed by transcribing the libraries in microtiter plates in the presence of Gemini-561. Indeed, a slight but progressive increase of R1 and R2 average fluorescence was observed with respect to R0 library (Fig. 3D). The resulting R2 enriched library was subjected to a third round of screening during which two sorting gates were used: a relaxed yet selective gate (population R3A, boxed in black on Fig. 3C) and a more stringent gate (population R3B, boxed in red on Fig. 3C) expected to contain aptamers in general more efficient than those exclusive to R3A. As a result of these slightly more stringent selections, both R3A and R3B libraries displayed a marked increase in fluorescence, with R3B performing better than R3A as expected (Fig. 3D). Since our objective was not necessarily to identify best performing aptamers but we rather wanted to preserve some sequence diversity, we decided to stop the process after this third round. Nevertheless, we were quite excited to see that both libraries displayed a higher fluorescence than the parental SRB-2 molecule, suggesting that improved mutants were likely present in these libraries. Sequences contained in both R3 libraries were indexed together with the starting library, and the three libraries (R0, R3A, and R3B) were finally sequenced on a MiSeq high-throughput sequencing platform.

Unsupervised bioinformatic sequencing data processing allows significant data reduction

Upon sequencing, reads were QC-filtered and those displaying mutations outside the initially randomized regions (i.e., UDI and P3) were discarded (see the overall analytical pipeline in Supplemental Fig. 1). Next, sequences sharing the same UDI were considered to originate from the same droplet and were clustered together. At this stage, only those sequences with an occurrence above an automatically computed threshold (Supplemental Fig. 2) were conserved. Indeed, and as more deeply explained elsewhere (Autour et al. 2019), UDI/sequence pairs with an occurrence below that threshold were likely point mutants raised from PCR or sequencing errors and were therefore no longer considered in the rest of the analysis. Then, counting the number of different UDIs associated with each P3 sequence allowed us to count the number of droplets

containing this sequence, so as to precisely compute the enrichment of each sequence. These sequences could then be ranked according to their occurrence (or enrichment) prior to functionally testing the most frequent ones as is done in the μIVC-seq approach we described before (Autour et al. 2019). However, this may leave a large fraction of sequences to test and only partly exploit the potential of the large data set collected upon sequencing.

To more deeply exploit sequencing data while reducing the information and restricting the functional assay to the most interesting sequences, we decided to use an artificial neural network algorithm to order and cluster the sequences in a planar grid by an unsupervised manner, considering therefore groups of sequences rather than individuals. In an attempt to define a set of relevant parameters to train the ANN, we first computed the minimum free energy (MFE) of each P3/L3 sequence using RNAfold from the ViennaRNA package (Lorenz et al. 2011) and found that the region tended to structure all the better as the selection stringency was increased (Fig. 4A). Indeed, while 2 to 6 bp tended to form in R0 library, this number was strongly biased toward the formation of 6 bp in libraries R3A and R3B (Fig. 4B), confirming the higher structuration of the selected molecules. Moreover, in the starting library (R0), most of the contacts are formed with L3 loop and only a very small fraction of molecule forms base pairs only in the P3 region (Fig. 4C). This somehow contrasts with molecules contained in R3A and B that rather tend to form the parental P3 stem and no interaction with L3 (Fig. 4C). These data not only confirm that, to be functional, the aptamer should adopt an SRB-2-like P3/L3 structure made of 6 bp long stem closed by a L3 region that should stay free of interaction with P3, but they also demonstrate the biological relevance of our analysis. Deeper sequence analysis did not reveal marked sequence preference other than an A_{31}–U_{42} base pair closing L3 and an overall A/U richness in R3A (Fig. 4D). Interestingly, in the better fit R3B library a subset of sequences tends to display an increased G/C content, especially a C_{31}–G_{42} base pair to which the parental sequence (G_{31}–C_{42}) does not conform (Fig. 4E). Altogether, these observations suggest that the establishment of 6 bp was a much stronger criterium than the stability of the formed stem (reflected by the MFE) and may therefore better drive the training of the ANN algorithm. A plausible explanation to this a priori surprising preference may be that P3 stem primarily acts as spacer between J2–3 and L3 unpaired regions and that distance (and perhaps the relative orientation) it imposes between both regions is more important that its intrinsic stability.

We next looked at the fitness landscape of R0, R3A, and R3B libraries. To do so, sequences were first organized in a two-dimension plan (x and y coordinates, Supplemental Fig. 1) using a self-organizing map (SOM), an artificial neural network (Kohonen 2013), prior to assigning the total occurrence of the group of sequence at each coordinate as the z value of the plot. The randomized sequence (i.e., nucleotides 27 to 31 and 42 to 46) of each molecule (note

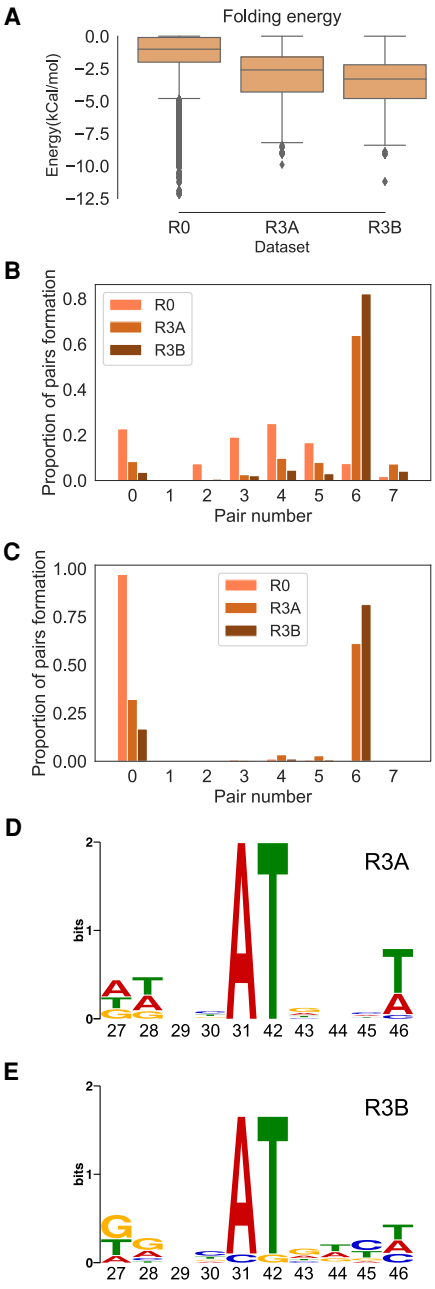


FIGURE 4. Characterization of the sequences contained in the starting and round three enriched libraries. (*A*) Folding energy of the sequences computed using RNAfold program from the ViennaRNA Package. (*B*) Base-pair formation in the P3–L3 stem–loop. The number of base pairs formed was extracted from the optimal secondary structure in dot-bracket notation generated by the RNAfold program of the ViennaRNA Package. (*C*) Base-pair formation in the P3 stem only. Here, any sequence displaying at least 1 bp involving L3 was set to 0. (*D*,*E*) Motif identified in the 50 sequences displaying the highest occurrence in R3A (*D*) or R3B (*E*) library. The motifs were generated using MEME algorithm (Bailey and Elkan 1994).

that we restricted the map to those sequences seen upon sequencing and that only 613,459 of the million expected sequences are represented) was first converted into a vector in which nucleotide identity was coded by its three-dimensional (3D) trajectory (TDT) as described in Lo et al. (2007). Briefly, each one of four nucleotides was assigned as one point in the 3D space, being the relative position of the nucleotides determined by the 3D coordinates of the four vertices of a regular tetrahedron. Using the sequence as the only information led to many (18 to 68) clusters (Supplemental Fig. 3A) leaving as many molecules to be functionally tested. As anticipated above, while adding the free energy of the molecule to the vector did not significantly reduced the number of peaks (Supplemental Fig. 3B), including in the vector both the free energy and the number of base pairs formed in the P3 region allowed us to cluster most of R3A and R3B sequences in two main clusters (Fig. 5A,B; Supplemental Fig. 3C), confirming the respective importance of each parameter. Remarkably, looking more closely at the content of each peak revealed that, whereas cluster 1 contains mainly A/U rich sequences highlighted above, those sequences contained in cluster 2 are rather rich in G/C (Fig. 5C). Consistently and as expected, the content of both clusters was rather homogeneous in terms of base pairs formed and energy, with cluster 2 displaying an overall lower free energy than cluster 1 (Supplemental Fig. 4). Finally, cluster 3 contained those sequences excluded from the two other clusters for which no composition preference was expected. Though our process starts with weights randomly assigned to each one of the map nodes during initiation (learning) step, it appeared to be quite robust since repeating it three times always allowed both clusters 1 and 2 to be identified (though at variable coordinates on the map due to the starting network randomization) with the similar sequence content (Supplemental Fig. 5).

Functional validation and identification of improved sequences

In order to functionally validate our bioinformatic clustering, we tested several sequences representative of each

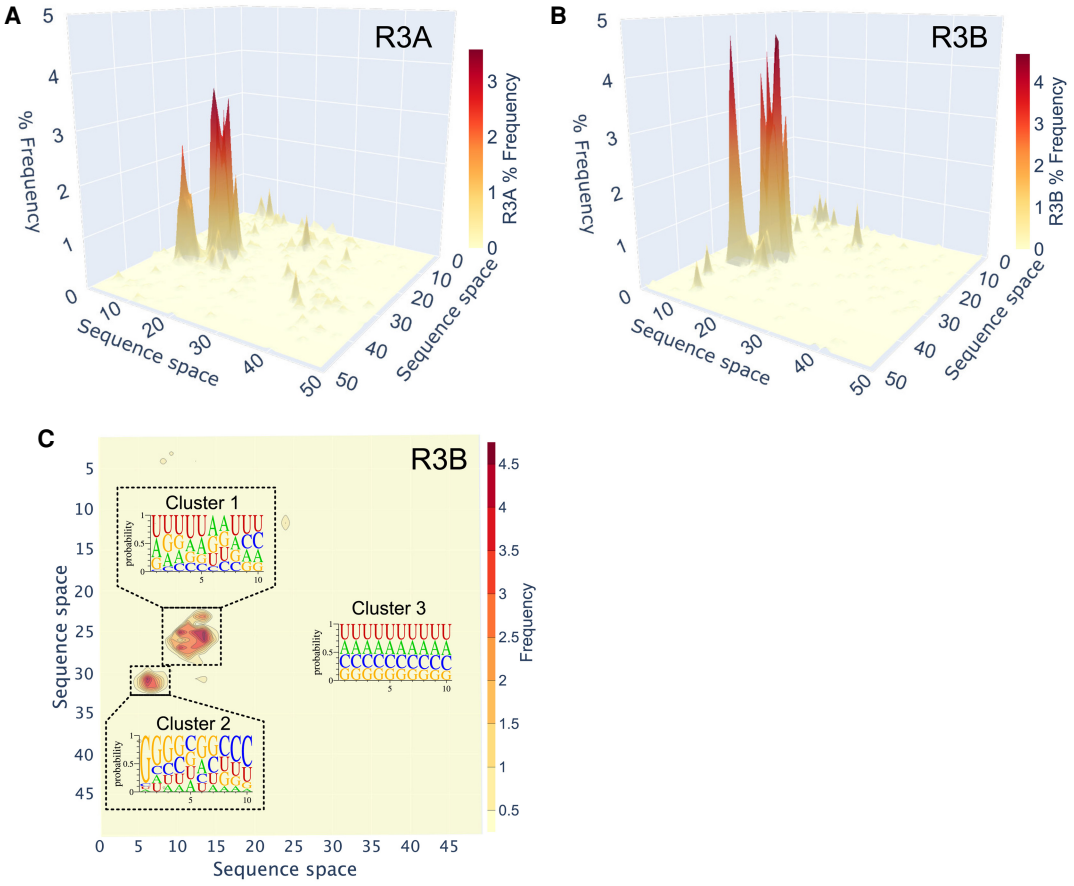
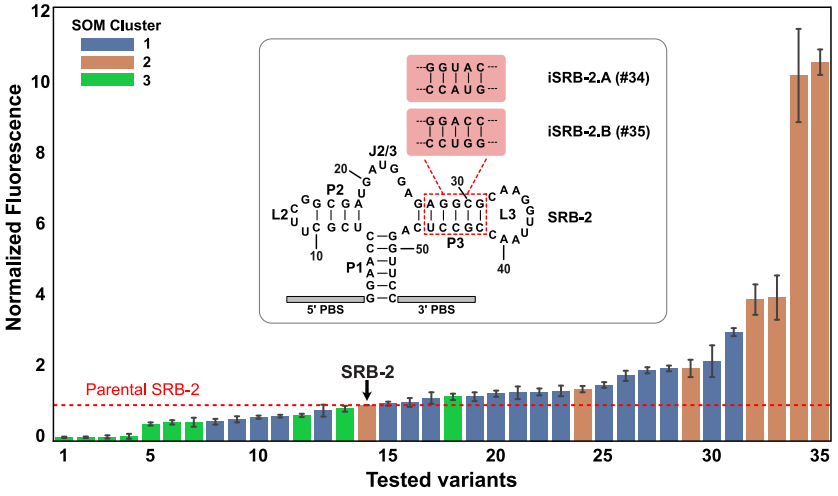


FIGURE 5. Fitness landscape of the analyzed sequences. (*A*) Fitness landscape of the R3A selection round generated from the map of the sequence space obtained upon sequence classification by the self-organizing map (SOM) algorithm. (*B*) Fitness landscape of the R3B selection round generated from the map of the sequence space obtained upon sequence classification by the SOM algorithm. (*C*) Contour plot of the R3B fitness landscape identifying three sequence clusters. Corresponding logo of the sequences contained in each cluster is also shown.

cluster. To do so, the sequences were in vitro transcribed in the presence of Gemini 561 and the emitted fluorescence was normalized to that of the parental SRB-2 aptamer (Fig. 6). Interestingly, we noticed important differences in the fluorescence of the complex formed with the different sequences. To rule out the possibility that this effect was due to variation in transcription efficiency, we quantified the amount of RNA produced at the end of measurement (Supplemental Fig. 6). Even though some templates (e.g., sequences 1 and 4) showed a higher transcription efficiency, the amount of RNA did not correlate with the emitted fluorescence, indicating that the observed variation was mainly linked to the capacity of the constructs to activate Gemini 561 fluorescence rather than their transcription efficiency. At first glance, no meaningful correlation could be observed between sequence occurrence and fluorescence emission when the sequences were directly ranked by their occurrence (Supplemental Fig. 7A), like this would be done during a μIVC-seq approach. However, clustering the sequences prior to ranking them allowed to nicely group and order together sequences with similar phenotypes (Supplemental Fig. 7B), the best variants originating from cluster 2. Indeed, the sequences coming from the G/C-rich cluster 2 tended to cluster together as those displaying the best function, an observation in good agreement with our earlier observation that G/C-rich sequences better accumulated in the best-fit R3B library (see above). Moreover, sequences from cluster 1 formed a distinct functional group that, even though displaying a lower fluorescence than cluster 2 variants, had a fluorescence significantly above that of sequences taken from cluster 3. Furthermore, among cluster 2 sequences, the two most represented variants (sequences 34 and 35 on Fig. 6) formed with Gemini-561 a complex an order of magnitude more fluorescent than the parental SRB-2. Based on the same nomenclature we used in our earlier work on Spinach (Autour et al. 2016) and Mango III (Trachman et al. 2019) aptamers, we named these aptamers iSRB-2A ($_{27}$GGUA C$_{31}$–$_{42}$GUACC$_{46}$) and iSRB-2B ($_{27}$GGACC$_{31}$–$_{42}$GGUCC$_{46}$). Excitingly, both sequences possessed the C$_{31}$–G$_{42}$ base pair predicted above and form two GC base pairs (G$_{27}$–C$_{46}$ and G$_{28}$–C$_{45}$) stabilizing the basis of P3 stem. Just looking at the free energy and the number of base pairs formed does not allow to distinguish these variants from the parental SRB-2 ($_{27}$AGGCG$_{31}$–$_{42}$CGCCU$_{46}$) sequence. Therefore, additional characterization will be needed in the future to decipher the origin of the 10-fold improvement observed. Yet, to get a first glimpse on how this $_{27}$GGNNC$_{31}$–$_{42}$ GNNCC$_{46}$ motif may be generic at eliciting efficient fluorescence emission, we tested the 36 possible sequence permutations (considering A–U, U–A, G–C, C–G, G.U, and U.G as possible base pairs established between nucleotides at positions 29 and 44, or 30 and 43). Interestingly, many of these sequences allowed an efficient fluorescence emission, especially when a purine was found at position 29 and a pyrimidine at position 30 or vice versa (Supplemental Fig. 8). Even though a deeper mechanistic characterization of the molecule will be required to explain this preference, these data not only confirmed our prediction that the $_{27}$GGNNC$_{31}$–$_{42}$ GNNCC$_{46}$ motif identified here indeed confers advantageous properties to the aptamer, but they also brought a strong functional validation of our unsupervised approach combining ultrahigh-throughput functional screening and unsupervised bioinformatic analysis.



Label	P3 Sequence	Label	P3 Sequence	Label	P3 Sequence
1	5'-TTTTT-3' // 5'-TATCT-3'	13	5'-AAATA-3' // 5'-TATTT-3'	24	5'-GGTTG-3' // 5'-CAACC-3'
2	5'-TTTTT-3' // 5'-ATTTT-3'	14	5'-ATAAG-3' // 5'-CTTAT-3'	25	5'-TATGA-3' // 5'-TCATA-3'
3	5'-GGTAC-3' // 5'-AAAAG-3'	SRB-2	5'-AGGCG-3' // 5'-CGCCT-3'	26	5'-GTACA-3' // 5'-TGTAT-3'
4	5'-GGTAC-3' // 5'-TTTAG-3'	15	5'-GTATA-3' // 5'-TATAC-3'	27	5'-ATACC-3' // 5'-GGTAT-3'
5	5'-TAAAG-3' // 5'-TTTTA-3'	16	5'-ATGCA-3' // 5'-TGTAT-3'	28	5'-GAACA-3' // 5'-TGTTT-3'
6	5'-TTATA-3' // 5'-TATGA-3'	17	5'-TTTCA-3' // 5'-TGAAA-3'	29	5'-GTACC-3' // 5'-GGTAT-3'
7	5'-CTACA-3' // 5'-TGTAA-3'	18	5'-GATAA-3' // 5'-TTATT-3'	30	5'-GTAGA-3' // 5'-TCTGC-3'
8	5'-TAATA-3' // 5'-TGTTA-3'	19	5'-AATGA-3' // 5'-TCATT-3'	31	5'-TGGCT-3' // 5'-AGTCA-3'
9	5'-GTATT-3' // 5'-AATAC-3'	20	5'-ATGTA-3' // 5'-TACAT-3'	32	5'-GCACT-3' // 5'-AGTGC-3'
10	5'-AAAAC-3' // 5'-GTTTT-3'	21	5'-GCATA-3' // 5'-TATGT-3'	33	5'-AGGGA-3' // 5'-TTCCT-3'
11	5'-GTTTG-3' // 5'-CAAAC-3'	22	5'-TATTG-3' // 5'-CAATA-3'	34	5'-GGTAC-3' // 5'-GTACC-3'
12	5'-TATTT-3' // 5'-AAATA-3'	23	5'-TTATG-3' // 5'-CATAA-3'	35	5'-GGACC-3' // 5'-GGTCC-3'

FIGURE 6. Functional analysis of SRB-2 mutants. Different sequences were selected from the self-organizing map (SOM) shown on Figure 5. Each construct was in vitro transcribed in the presence of 500 nM of Gemini-561 and the fluorescence monitored over time. The fluorescence apparition rate was computed and normalized to that of the parental SRB-2 aptamer. The values are the mean of $n = 3$ independent experiments. The error bars correspond to ±1 SD. The bars are color-coded with respect to the SOM cluster from which the sequence was originally selected. The P3 sequence of each variant is given in the table *under* the plot. Sequences of both strands of the stem were concatenated into a single line.

DISCUSSION

For a long time, the analysis of nucleic acids isolated by SELEX and other in vitro selection procedures was limited to the few tens of sequences that are generally the most represented ones in the final population. However, technological advances like NGS now allow the whole selection process to be characterized at once and is commonly used to assist hits identification (Nguyen Quang et al. 2016; Kinghorn et al. 2017; Yokobayashi 2019). The global view offered by such analyses makes it possible to rapidly identify variants of interest, even though they are underrepresented at the end of the selection process (Nguyen Quang et al. 2018; Autour et al. 2019). Yet, these approaches may still require significant manual analyses and may be fastidious. In the present work, we further increased the automation level by using NGS in tandem with an artificial neural network algorithm. An important point in the development of this new methodology was the proper encoding of nucleotide in a vector format allowing the SOM algorithm to properly handle sequences. Original attempts to encode nucleotides composing a sequence in a binary format failed at producing convincing clustering (data not shown), driving us to explore alternative encoding formats. Eventually, we found that using the TDT codification, in which each nucleotide of a given sequence is encoded by the coordinates of a regular tetrahedron vertices (Lo et al. 2007) prior to being concatenated in its 3D sequence trajectory, gave the best results. Yet, this was not sufficient to get tight clustering and required additional information about energy and base-pairing to be included as well. Note that all this information was directly collected by the algorithm, therefore enabling the use of a unique pipeline incorporating all the functionalities (Supplemental Fig. 1). Using this pipeline, we managed to reduce the overall sequence information of a functional screening process initiated with ~1 million different sequences down to 2 clusters of functional sequences, of which one contained at least several major sequences endowed with a significantly improved (~10 times better) function. Therefore, we are confident that applying this strategy to the analysis of other selection processes may further improve the discovery rates of highly efficient molecules.

In the work described herein, we used this unsupervised bioinformatic analysis in tandem with the ultrahigh-throughput microfluidic-assisted functional screening we originally named μIVC (Ryckelynck et al. 2015). μIVC has previously demonstrated its great efficiency at selecting RNAs endowed with functions (e.g., catalysis, fluorogen lighting-up) involving a phenotype (e.g., substrate cleavage, fluorescence emission) that physically dissociates from the RNA. Indeed, by confining biological reactions into picoliter volume droplets, this technology allows to functionally and accurately assay millions of molecules in a single experiment (Autour and Ryckelynck 2017; Bouhedda et al. 2018). Though robust, this technology initially suffered from the same limitation than other in vitro selection technologies, that is, the analysis of only a small subset of the most abundant sequences. We recently reported on a first level of improvement by using μIVC in tandem with NGS (a method called μIVC-seq) to rapidly identify optimized communication modules during the development of small molecule biosensors (Autour et al. 2019). However, the computer-assisted analysis of the selected sequences still required extensive intervention of the experimenter that are no longer needed in the present format of μIVC pipeline we propose to call μIVC-Useq for μIVC coupled with Unsupervised sequence analysis. As shown in this work, this new approach nicely complements and reinforces the set of existing technologies available to characterize RNA structure/function relationship.

We choose the light-up RNA aptamer SRB-2 (Holeman et al. 1998) as model system to set-up and demonstrate the efficiency of μIVC-Useq. This aptamer has been extensively used for the development of new efficient RNA imaging tools either by directly using the RNA as is (Sunbul and Jaschke 2013, 2018) or by further evolving it to recognize other fluorogens (Bouhedda et al. 2020). However, despite this strong interest and the great perspectives of this aptamer, the exact tridimensional structure of the molecule remains unknown, leaving important questions on its intimate working mechanism and capacity to sustain engineering largely unanswered. Indeed, whereas P1 stem (Bouhedda et al. 2020) and P2/L2 stem–loop (Holeman et al. 1998) were successfully modified without compromising SRB-2 sulforhodamine B-binding capacity, less was known about the P3/L3 stem–loop except that, together with J2/3, it encompasses the sulforhodamine B-binding site. Therefore, to shed further light on this element we prepared a mutant library in which 5 of the 6 bp of P3 were randomized. Performing three rounds of μIVC screening on this library followed by NGS sequencing and the use of our unsupervised bioinformatic pipeline allowed us to confirm the existence of the up-to-now putative P3 stem as well as to draw several new important conclusions on the P3/L3 region of SRB-2. First, all the functional molecules possess a 6 bp long P3 helix. Therefore, the length of P3 should play an important role in the proper structuration of sulforhodamine B-binding site since, in our experiment, the RNA had the possibility to acquire shorter stems, which it did not. Second, the sequence of the stem should not allow interaction with L3 to take place as sequences able to establish such contact, though present in the starting library, were strongly counter selected. This observation further supports a key function of L3 loop in the recognition of the dye. Third, though a significant variety of sequences able to form the required stem was tolerated by the molecule, they lead to a wide range of phenotypes spanning an order of magnitude.

Therefore, even though they are conservative and would be predicted as being optimal by a computer-assisted RNA folding prediction software, not all sequences are tolerated the same way. This is typically exemplified in our study by iSRB-2A and B mutants that both have the same overall free energy as the parental SRB-2 but display a 10-fold better capacity to activate Gemini-561 fluorescence. The exact mechanism by which these mutations act would require a dedicated study that is out of the scope of the present one. Nevertheless, at this stage, one may imagine different scenarios such as the direct promotion of extra stabilizing tertiary contacts with the loop elements or, on the opposite side, a more indirect effect by which the mutant sequences would prevent/limit the formation of undesirable alternative folding prone to reduce the fraction of molecules competent to bind the dye. Whatever the mechanism at work, it is extremely unlikely that any rational design approach would have been able to predict such sequences, further reinforcing the need of high-throughput technologies like μIVC-seq and now μIVC-Useq to properly assist the design and engineering of RNA molecules.

MATERIALS AND METHODS

Library design

The template sequence was designed on the basis of SRB-2 aptamer (Holeman et al. 1998). The P3 stem was randomized 5′-GGAACCTCGCTTCGGCGATGATGGAG**NNNNN**CAAGGTTAAC**NNNNN**CAGGTTCC-3′ and flanked with a 5′ (5′-GGGAGACAGCTAGAGTAC-3′) and a 3′(5′-GTACACTGTGCTCGTGTC-3′) constant regions to yield *SRB-2 P3N10-ext* template (Supplemental Table S2).

DNA amplification and barcoding

To allow the transcription and the identification of each variant of the library, a T7 RNA polymerase promoter and a UDI random barcode (Autour et al. 2019) were appended to the 5′ end of *SRB-2 P3N10-ext* template by PCR amplification. To do so, 1 pmol of the *SRB-2 P3N10-ext* template library was introduced in 100 µL of PCR mixture containing 0.5 µM of primer 2.A (5′-*TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG*aaNNNNNNNNNNNNNNNNNNaaTATcTAATACGACTCACTATAGGGAGACAGCTAGAGTAC-3′) and 2.B (5′-GTACACTGTGCTCGTGTC-3′), 0.2 mM each dNTPs, 1X Evagreen (Biotium), 2 U of Q5 DNA polymerase (New England Biolabs) and the corresponding buffer at the recommended concentration. An amount of 20 µL of this mixture was introduced into a qPCR machine (CFX-96, Bio-Rad) and was thermocycled starting with an initial step of denaturation of 30 sec at 95°C, followed by 40 cycles of 5 sec at 95°C, 30 sec at 60°C and 30 sec at 72°C. Upon determination of the threshold cycle (Ct), the remaining 80 µL of mixture were thermocycled at Ct+2 cycles. PCR products were purified on gel by the Wizard SV Gel and PCR Clean-up System kit (Promega) prior to being quantified by NanoDrop.

Then, a second PCR was performed to amplify the barcoded sequences. 0.1 pmol of the former PCR products was introduced in 100 µL of PCR mixture containing 0.5 µM of each primer 3.A (5′-TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG-3′) and 1.B (5′-GTACACTGTGCTCGTGTC-3′), 0.2 mM each dNTPs, 1X Evagreen (Biotium), 2 U of Q5 DNA polymerase (New England Biolabs) and the corresponding buffer at the recommended concentration. An amount of 20 µL of this mix was introduced into a qPCR machine (CFX-96, Bio-Rad) as above and, upon determination of the threshold cycle (Ct), the remaining 80 µL of mixture was thermocycled at Ct+2 cycles. PCR products were purified by the Wizard SV Gel and PCR Clean-up System kit (Promega).

Droplet-based microfluidic screening

Microfluidic chips were fabricated in polydimethylsiloxane (PDMS) as described in Ryckelynck et al. (2015).

Droplet digital PCR

DNA mutant libraries were diluted in 200 µg/mL yeast total RNA solution (Ambion) to obtain the desired occupancy of droplets. An amount of 1 µL of this dilution was then introduced in 100 µL of PCR mixture containing 0.5 µM of each primer 3.A (5′-TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG-3′) and 1.B (5′-GTACACTGTGCTCGTGTC-3′), 0.2 mM each dNTPs, 20 µM coumarin acetate (Sigma-Aldrich), 0.1% Pluronic F68 (Sigma-Aldrich), 2 U of Q5 DNA polymerase (New England Biolabs) and the corresponding buffer at the recommended concentration. The mixture was loaded in a length of PTFE tubing and infused into a droplet generator microfluidic chip where it was dispersed in 2.5 pL droplets (production rate of ∼12,000 droplets/s) carried by Novec 7500 fluorinated oil (3M) supplemented with 3% of a fluorosurfactant (proprietary synthesis). Droplet production frequency was monitored in real time using an optical device and software developed by the team (Ryckelynck et al. 2015) and used to determine droplet volume. 2.5 pL droplets were generated by adjusting pumps flowrates (MFCS, Fluigent). The emulsion was collected in 0.2 mL tubes and subjected to an initial denaturation step of 2 min at 98°C followed by 30 cycles of: 10 sec at 98°C, 30 sec at 55°C, 30 sec at 72°C. Droplets were then reinjected into a droplet fusion microfluidic device.

Droplet fusion

PCR droplets were reinjected and spaced into a fusion device at a rate of ∼1500 droplets/sec. Each PCR droplet was then synchronized with a 16 pL IVT droplet containing 2 mM each NTP (Larova), 25 mM $MgCl_2$, 44 mM Tris-HCl pH 8.0 (at 25°C), 5 mM DTT, 1 mM Spermidine, 0.1% of Pluronic F68 (Sigma-Aldrich), 1 µg of pyrophosphatase (Roche), 500 nM Gemini-561, 1 µM coumarin acetate (Sigma-Aldrich) and 17.5 µg/mL T7 RNA polymerase (purified in the laboratory). IVT mixture was loaded in a length of PTFE tubing and kept on ice during all the experiment. PCR droplets were spaced and IVT droplets produced using a dedicated stream of Novec 7500 fluorinated oil (3M) supplemented with 2% (w/w) of fluorosurfactant. Flowrates (MFCS, Fluigent) were adjusted to generate 16 pL IVT droplets and maximize synchronization of 1 PCR droplet with 1 IVT droplet. Pairs of droplets

were then fused with an AC field (400 V at 30 kHz) and the resulting emulsion was collected off-chip and incubated for 2 h at 37°C.

Droplet sorting

The emulsion was finally reinjected into an analysis and sorting microfluidic device at a frequency of ~150 droplets/sec and spaced with a stream of surfactant-free Novec 7500 fluorinated oil (3M). The orange fluorescence (Gemini-561 in complex with an aptamer) of each droplet was analyzed and the most orange fluorescent droplets were sorted (from 6.7% to 0.18% depending on the round of μIVC, Supplemental Table S1). The gated droplets were deflected into collecting channel by applying an AC fields (1000 V at 30 kHz) and collected into a 1.5 mL tube. Sorted droplets were recovered from the collection tubing by flushing 200 μL of HFE fluorinated oil (3M). An amount of 100 μL of 1H, 1H, 2H, 2H-perfluoro-1-octanol (Sigma-Aldrich) and 200 μL of 200 μg/mL yeast total RNA solution (Ambion) were then added and the droplets broken by vortexing the mixture. DNA-containing aqueous phase was then transferred into a classical Eppendorf tube.

Amplification of sorted DNA

An amount of 2 μL of aqueous phase obtained upon droplet breaking were introduced in 100 μL of PCR mixture containing primers 2.A and 1.B as above (see "DNA amplification and barcoding" section) to reset the UDI carried by the DNA. This was essential to preserve the quantitative capacity of the method over successive μIVC rounds (Autour et al. 2019). The DNA was treated as above and an aliquot of purified PCR product was subjected to a second PCR using primers 3.A and 1.B.

Libraries indexing for high-throughput sequencing

The starting library and those obtained upon each round of screening were indexed using Nextera technology (Illumina). First, a Nextera-compatible sequence was appended to the 3′ end of each gene. To do so, 2 μL of aqueous phase obtained upon droplet breaking were introduced in 100 μL of PCR mixture containing 0.5 μM of each primer 3.A (5′-TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG-3′) and 2.B (5′-GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGGTACACTGTGCTCGTGTC-3′), 0.2 mM each dNTPs, 1X Evagreen (Biotium), 2 U of Q5 DNA polymerase (New England Biolabs) and the corresponding buffer at the recommended concentration. An amount of 20 μL of this mix were introduced into a qPCR machine (CFX-96, Bio-Rad) as above and, upon determination of the threshold cycle (Ct), the remaining 80 μL of mixture were thermocycled at Ct+2 cycles. PCR products were purified by the Wizard SV Gel and PCR Clean-up System kit (Promega).

A second PCR was then performed to add Illumina indexes both at the 5′ and 3′ ends of each recovered DNA molecules. An amount of 0.1 pmol of the first PCR product was introduced in 100 μL of PCR mixture containing 0.5 μM of each Nextera Illumina primer (N7 and N5; a different pair for each library to index), 0.2 mM each dNTPs, 1X Evagreen (Biotium), 2 U of Q5 DNA polymerase (New England Biolabs) and the corresponding buffer at the recommended concentration. 20 μL of this mix were introduced into a qPCR machine (CFX-96, Bio-Rad) as above and, upon determination of the threshold cycle (Ct), the remaining 80 μL of mixture were thermocycled at Ct+2 cycles. PCR products were purified by the Wizard SV Gel and PCR Clean-up System kit (Promega). Libraries were finally loaded on a V3-150 chip (Illumina) and analyzed on a MiSeq sequencing platform (Illumina).

Bioinformatic sequence analysis

Sequencing data were analyzed using a custom Python bioinformatic pipeline in 10 main steps (Supplemental Fig. 1). First, fastq files were parsed using the Biopython library and only reads with a Q-score ≥ 30 were conserved for the rest of the analysis (step 1). Then, UDI and 10-mer randomized regions were extracted from each read (step 2). An UDI occurrence cut-off was automatically set for each library (Supplemental Fig. 2). Sequences with an occurrence below that threshold were likely mutants (raised from PCR or sequencing errors) and were no longer considered for the rest of the analysis (step 3). Moreover, sequences displaying mutations outsides of randomized regions (i.e., the UDI and the P3 stem) were also filtered out. Next, identical sequences with different UDI and the expected length were clustered together, and their occurrence measured (step 4) while the 10-mer randomized sequences of each selection round were isolated in parallel (step 5). Sequences of the P3/L3 stem–loop regions were used to compute the MFE and determine the number of base pairs formed using RNAup from the RNAlib python library of the ViennaRNA Package (step 6). Nucleotide sequences were codified in TDT vectors as described in Lo et al. (2007) and, in some analyses, MFE and the number of base pairs formed were added to the sequence TDT vector (step 7). Using the SOMPY python library, a SOM of the sequence space was trained using the sequences TDT vectors generated in step 7, eventually appending MFE or the MFE and the nucleotide pair number formation to the vector (step 8). A grid of 50 × 50 neurons with randomly weights generation was selected to represent the sequence space. To train the model we used a rough train with 40 iterations with a radius of 10 followed by 80 fine-tuned train iterations with a radius of 4. Next, a fitness landscape was constructed from the nodes grid generated by the SOM algorithm (step 9). For each node a fitness value (Z axis of the fitness landscape) was calculated by adding the sum of the occurrence frequency of all the sequences present on that node. Finally, neighboring nodes sharing a high fitness were clustered together in view of further analyzing their sequence content and features.

Functional validation of selected sequences

For each tested sequence, a template oligonucleotide was chemically synthetized by IDT (Integrated DNA Technologies). An amount of 0.1 pmol of template was then added to 20 μL of PCR mixture containing 0.5 μM of each primer 1.A (5′-CTTTAATACGACTCACTATAGGGAGACAGCTAGAGTAC-3′, adding the T7 promotor) and 1.B (5′-GTACACTGTGCTCGTGTC-3′), 0.2 mM each dNTPs, 2 U of Q5 DNA polymerase (New England Biolabs) and the corresponding buffer at the recommended concentration. The PCR mixtures were thermocycled for 25 cycles starting with an initial step of denaturation of 30 sec at 95°C

followed by 40 cycles of 5 sec at 95°C, 30 sec at 60°C and 30 sec at 72°C. PCR products were then purified by the Wizard SV Gel and PCR Clean-up System kit (Promega) and quantified by NanoDrop.

A total of 40 ng of purified DNA were then introduced in 40 µL of in vitro transcription mixture containing 2 mM each NTP (Larova), 25 mM $MgCl_2$, 44 mM Tris-HCl pH 8.0 (at 25°C), 5 mM DTT, 1 mM Spermidine, 1 µg of pyrophosphatase (Roche), 17.5 µg/mL T7 RNA polymerase (purified in the laboratory) and 500 nM of Gemini-561. This mixture was then incubated at 37°C in a real-time thermocycler (Stratagene Mx3005P, Agilent Technologies) and the fluorescence of the reaction was monitored for 2 h (ex/em 575 nm/602 nm). Note that library enrichment monitoring was performed following exactly the same procedure.

SUPPLEMENTAL MATERIAL

Supplemental material is available for this article.

ACKNOWLEDGMENTS

We thank Sandrine Koechler and Abdelmalek Alioua from the IBMP Gene Expression Analysis facility (funded by LabEx NetRNA) for technical assistance with high-throughput sequencing. This work received financial support from the University of Strasbourg Institute of Advanced Study (USIAS, program Translatomix) and Agence Nationale de la Recherche (BrightRiboProbes, ANR-16-CE11-0010-01) and SATT Conectus (prematuration program LUNA). This work of the Interdisciplinary Thematic Institute "IMCBio," as part of the ITI 2021-2028 program of the University of Strasbourg, CNRS and Inserm, was supported by IdEx Unistra (ANR-10-IDEX-0002), the SFRI-STRAT'US project, and EUR IMCBio (ANR-17-EURE-0023) under the framework of the French Investments for the Future Program. It was also supported by the Centre National de la Recherche Strasbourg. Finally, it also received support from the Initiative of Excellence (IdEx) of the Université de Strasbourg.

Received August 11, 2020; accepted May 1, 2021.

REFERENCES

Autour A, Ryckelynck M. 2017. Ultrahigh-throughput improvement and discovery of enzymes using droplet-based microfluidic screening. *Micromachines (Basel)* **8:** 128. doi:10.3390/mi8040128

Autour A, Westhof E, Ryckelynck M. 2016. iSpinach: a fluorogenic RNA aptamer optimized for in vitro applications. *Nucleic Acids Res* **44:** 2491–2500. doi:10.1093/nar/gkw083

Autour A, Jeng SCY, Cawte AD, Abdolahzadeh A, Galli A, Panchapakesan SSS, Rueda D, Ryckelynck M, Unrau PJ. 2018. Fluorogenic RNA Mango aptamers for imaging small non-coding RNAs in mammalian cells. *Nat Commun* **9:** 656. doi:10.1038/s41467-018-02993-8

Autour A, Bouhedda F, Cubi R, Ryckelynck M. 2019. Optimization of fluorogenic RNA-based biosensors using droplet-based microfluidic ultrahigh-throughput screening. *Methods* **161:** 46–53. doi:10.1016/j.ymeth.2019.03.015

Bailey TL, Elkan C. 1994. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc Int Conf Intell Syst Mol Biol* **2:** 28–36.

Baret JC, Miller OJ, Taly V, Ryckelynck M, El-Harrak A, Frenz L, Rick C, Samuels ML, Hutchison JB, Agresti JJ, et al. 2009. Fluorescence-activated droplet sorting (FADS): efficient microfluidic cell sorting based on enzymatic activity. *Lab Chip* **9:** 1850–1858. doi:10.1039/b902504a

Barrick JE, Corbino KA, Winkler WC, Nahvi A, Mandal M, Collins J, Lee M, Roth A, Sudarsan N, Jona I, et al. 2004. New RNA motifs suggest an expanded scope for riboswitches in bacterial genetic control. *Proc Natl Acad Sci* **101:** 6421–6426. doi:10.1073/pnas.0308014101

Bartel DP, Zapp ML, Green MR, Szostak JW. 1991. HIV-1 Rev regulation involves recognition of non-Watson-Crick base pairs in viral RNA. *Cell* **67:** 529–536. doi:10.1016/0092-8674(91)90527-6

Bouhedda F, Autour A, Ryckelynck M. 2018. Light-up RNA aptamers and their cognate fluorogens: from their development to their applications. *Int J Mol Sci* **19:** 44. doi:10.3390/ijms19010044

Bouhedda F, Fam KT, Collot M, Autour A, Marzi S, Klymchenko A, Ryckelynck M. 2020. A dimerization-based fluorogenic dye-aptamer module for RNA imaging in live cells. *Nat Chem Biol* **16:** 69–76. doi:10.1038/s41589-019-0381-8

Carothers JM, Oestreich SC, Davis JH, Szostak JW. 2004. Informational complexity and functional activity of RNA structures. *J Am Chem Soc* **126:** 5130–5137. doi:10.1021/ja031504a

Denny SK, Greenleaf WJ. 2019. Linking RNA sequence, structure, and function on massively parallel high-throughput sequencers. *Cold Spring Harb Perspect Biol* **11:** a032300. doi:10.1101/cshperspect.a032300

Ellington AD, Szostak JW. 1990. In vitro selection of RNA molecules that bind specific ligands. *Nature* **346:** 818–822. doi:10.1038/346818a0

Ennifar E. 2016. *Nucleic acid crystallography: methods and protocols.* Humana Press, New York.

Filonov GS, Moon JD, Svensen N, Jaffrey SR. 2014. Broccoli: rapid selection of an RNA mimic of green fluorescent protein by fluorescence-based selection and directed evolution. *J Am Chem Soc* **136:** 16299–16308. doi:10.1021/ja508478x

Holeman LA, Robinson SL, Szostak JW, Wilson C. 1998. Isolation and characterization of fluorophore-binding RNA aptamers. *Fold Des* **3:** 423–431. doi:10.1016/S1359-0278(98)00059-5

Holtze C, Rowat AC, Agresti JJ, Hutchison JB, Angile FE, Schmitz CHJ, Koster S, Duan H, Humphry KJ, Scanga RA, et al. 2008. Biocompatible surfactants for water-in-fluorocarbon emulsions. *Lab Chip* **8:** 1632–1639. doi:10.1039/b806706f

Husser C, Dentz N, Ryckelynck M. 2021. Structure-switching RNAs: from gene expression regulation to small molecule detection. *Small Struct* **2:** 2000132. doi:10.1002/sstr.202000132

Islam S, Zeisel A, Joost S, La Manno G, Zajac P, Kasper M, Lonnerberg P, Linnarsson S. 2014. Quantitative single-cell RNA-seq with unique molecular identifiers. *Nat Methods* **11:** 163–166. doi:10.1038/nmeth.2772

Ketterer S, Fuchs D, Weber W, Meier M. 2015. Systematic reconstruction of binding and stability landscapes of the fluorogenic aptamer spinach. *Nucleic Acids Res* **43:** 9564–9572. doi:10.1093/nar/gkv944

Kinghorn AB, Fraser LA, Lang S, Shiu SCC, Tanner JA. 2017. Aptamer bioinformatics. *Int J Mol Sci* **18**.

Kohonen T. 2013. Essentials of the self-organizing map. *Neural Netw* **37:** 52–65. doi:10.1016/j.neunet.2012.09.018

Lo NW, Chang HT, Xiao SW, Li CH, Kuo CJ. 2007. Global visualization and comparison of DNA sequences by use of three-dimensional trajectories. *J Inf Sci Eng* **23:** 1723–1736.

Lorenz R, Bernhart SH, Honer Zu Siederdissen C, Tafer H, Flamm C, Stadler PF, Hofacker IL. 2011. ViennaRNA Package 2.0. *Algorithms Mol Biol* **6:** 26. doi:10.1186/1748-7188-6-26

Mazutis L, Araghi AF, Miller OJ, Baret JC, Frenz L, Janoshazi A, Taly V, Miller BJ, Hutchison JB, Link D, et al. 2009. Droplet-based microfluidic systems for high-throughput single DNA molecule isothermal amplification and analysis. *Anal Chem* **81:** 4813–4821. doi:10.1021/ac900403z

Nguyen Quang N, Perret G, Duconge F. 2016. Applications of high-throughput sequencing for in vitro selection and characterization of aptamers. *Pharmaceuticals (Basel)* **9:** 76. doi:10.3390/ph9040076

Nguyen Quang N, Bouvier C, Henriques A, Lelandais B, Duconge F. 2018. Time-lapse imaging of molecular evolution by high-throughput sequencing. *Nucleic Acids Res* **46:** 7480–7494. doi:10.1093/nar/gky583

Pavlova N, Kaloudas D, Penchovsky R. 2019. Riboswitch distribution, structure, and function in bacteria. *Gene* **708:** 38–48. doi:10.1016/j.gene.2019.05.036

Ryckelynck M, Baudrey S, Rick C, Marin A, Coldren F, Westhof E, Griffiths AD. 2015. Using droplet-based microfluidics to improve the catalytic properties of RNA under multiple-turnover conditions. *RNA* **21:** 458–469. doi:10.1261/rna.048033.114

Schneider TD, Stormo GD, Gold L, Ehrenfeucht A. 1986. Information content of binding sites on nucleotide sequences. *J Mol Biol* **188:** 415–431. doi:10.1016/0022-2836(86)90165-8

Sherlock ME, Breaker RR. 2020. Former orphan riboswitches reveal unexplored areas of bacterial metabolism, signaling, and gene control processes. *RNA* **26:** 675–693. doi:10.1261/rna.074997.120

Sunbul M, Jaschke A. 2013. Contact-mediated quenching for RNA imaging in bacteria with a fluorophore-binding aptamer. *Angew Chem Int Ed Engl* **52:** 13401–13404. doi:10.1002/anie.201306622

Sunbul M, Jaschke A. 2018. SRB-2: a promiscuous rainbow aptamer for live-cell RNA imaging. *Nucleic Acids Res* **46:** e110. doi:10.1093/nar/gky543

Trachman RJ, Autour A, Jeng SCY, Abdolahzadeh A, Andreoni A, Cojocaru R, Garipov R, Dolgosheina EV, Knutson JR, Ryckelynck M, et al. 2019. Structure and functional reselection of the Mango-III fluorogenic RNA aptamer. *Nat Chem Biol* **15:** 472–479. doi:10.1038/s41589-019-0267-9

Tuerk C, Gold L. 1990. Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. *Science* **249:** 505–510. doi:10.1126/science.2200121

Turner DH, Mathews DH. 2016. *RNA structure determination: methods and protocols*. Humana Press, New York.

Wang J, Gong Q, Maheshwari N, Eisenstein M, Arcila ML, Kosik KS, Soh HT. 2014. Particle display: a quantitative screening method for generating high-affinity aptamers. *Angew Chem Int Ed Engl* **53:** 4796–4801. doi:10.1002/anie.201309334

Weinberg Z, Wang JX, Bogue J, Yang J, Corbino K, Moy RH, Breaker RR. 2010. Comparative genomics reveals 104 candidate structured RNAs from bacteria, archaea, and their metagenomes. *Genome Biol* **11:** R31. doi:10.1186/gb-2010-11-3-r31

Weinberg Z, Lunse CE, Corbino KA, Ames TD, Nelson JW, Roth A, Perkins KR, Sherlock ME, Breaker RR. 2017. Detection of 224 candidate structured RNAs by comparative analysis of specific subsets of intergenic regions. *Nucleic Acids Res* **45:** 10811–10823. doi:10.1093/nar/gkx699

Winkler WC, Breaker RR. 2005. Regulation of bacterial gene expression by riboswitches. *Annu Rev Microbiol* **59:** 487–517. doi:10.1146/annurev.micro.59.030804.121336

Yokobayashi Y. 2019. Applications of high-throughput sequencing to analyze and engineer ribozymes. *Methods* **161:** 41–45. doi:10.1016/j.ymeth.2019.02.001

Zadeh JN, Steenberg CD, Bois JS, Wolfe BR, Pierce MB, Khan AR, Dirks RM, Pierce NA. 2011. NUPACK: analysis and design of nucleic acid systems. *J Comput Chem* **32:** 170–173. doi:10.1002/jcc.21596

Zuker M. 2003. Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res* **31:** 3406–3415. doi:10.1093/nar/gkg595

μIVC-Useq: a microfluidic-assisted high-throughput functional screening in tandem with next-generation sequencing and artificial neural network to rapidly characterize RNA molecules

Roger Cubi, Farah Bouhedda, Mayeul Collot, et al.

RNA 2021 27: 841-853 originally published online May 5, 2021
Access the most recent version at doi:10.1261/rna.077586.120