Closing the loop on crowd-sourced science

James M. Robson[1,2] and Alexander A. Green[1,2,3]*

[1]Department of Biomedical Engineering, Boston University, Boston, MA 02215, USA

[2]Biological Design Center, Boston University, Boston, MA 02215, USA

[3]Molecular Biology, Cell Biology & Biochemistry Program, Graduate School of Arts and Sciences, Boston University, Boston, MA 02215, USA

*Corresponding author, aagreen@bu.edu

In the age of smartphones, tablets, and personal computers, information communication technologies have changed the way the world thinks, works, and functions. Digital technologies have influenced how people communicate with one another and how knowledge is shared. The dissemination of scientific knowledge has been revolutionized by the internet, enabling researchers to share their findings and data directly with followers outside the scientific community. While online sharing has facilitated scientific collaboration and aided in the distribution of scientific results, it has been far more challenging to bring the outside community into the process of scientific discovery and hypothesis testing to take advantage of crowdsourced insights. In PNAS, Andreasson et al. (1) present a platform for iterative hypothesis generation and high-throughput characterization that enables large-scale, videogame-based crowdsourcing of tens of thousands of RNA-based sensor designs. They find that bringing the wisdom of an online community into the discovery process yields unanticipated RNA device architectures along with sensors that operate near the thermodynamic optimum.

RNA is an ideal polymer for biomolecular design: it can code for genetic information, fold into intricate structures known as aptamers to capture ligand molecules, catalyze chemical reactions, and it plays a role in nearly every biological process in living cells. Moreover, the pairing relationships of the RNA bases, where A binds to U and G binds to C, provide a means to direct how RNA molecules fold and how they interact with other transcripts. RNA therefore provides a promising way to construct new regulatory elements based on its diversity of functions and our growing understanding of sequence-function relationships. Indeed, a diverse assortment of engineered RNA-based switches have been described from riboswitches that bind to different metabolites and proteins to regulate gene expression (2–4) to riboregulators that can sense pathogen RNAs (5) and perform biomolecular computations (6).

While RNA switches hold great promise in many applications, their design remains challenging due to sequence-specific interactions, the difficulties in predicting the energies of 3D conformations, and non-canonical base pairs that have not been extensively characterized. The fundamental rules that determine RNA switch performance remain poorly understood. Without rules to explain relationships between RNA sequence, structure, and behavior, rapid generation of RNA elements for biological control systems remains a challenge. Sequence design using software packages like NUPACK (7) and ViennaRNA (8) allow for the generation of thousands of potential sensors, but many RNA designs do not perform well when tested experimentally (9), making their performance difficult to predict, and thus, slowing development.

In PNAS, Andreasson et al. (1) extend the iterative process of RNA design, synthesis, and testing, describing a pipeline for “citizen scientists” to make, validate, and modify hypotheses at scale, thereby accelerating design advances for RNA switches. Thousands of sensor designs are crowdsourced through

Eterna (10), a massive open laboratory and discovery game, enabling the authors to overcome the limitations of existing computational RNA software design packages by leveraging both human insight and experimental testing (Fig. 1). The first community design challenge focused on implementing an RNA switch to detect the cellular metabolite flavin mononucleotide (FMN). Eterna players were tasked with engineering the RNA switch to form the MS2 aptamer RNA structure after binding of FMN. The MS2 aptamer can in turn capture a fluorescent protein ligand, enabling visualization under a microscope. Using repurposed Illumina sequencing chips, the player-contrived RNA switches were evaluated at high throughput, providing quantitative RNA characterization data. High-performing player-designed switches from the Eterna community were compared to Ribologic (11), an automated algorithm for designing RNA molecules that are predicted to change their secondary structure in response to interactions with other molecules. Computer-generated designs via Ribologic were found to exhibit lower activation ratios than the final Eterna player designs after iterative refinement.

The generality of the crowdsourced design approach was tested by developing RNA sensors for small molecules with different aptamer reporters. In particular, Andreasson et al. (1) demonstrate RNA switches with malachite green and Spinach aptamer reporters (12, 13), which bind dye molecules to generate fluorescent signals. They show that when the total number of tested designs is reduced, the Eterna community players generate responsive switches even in cases when the Ribologic algorithm fails. The authors go on to demonstrate that the optimal switch design, regardless of the type of riboswitch, requires testing not only multiple switch designs, but also multiple switch architectures. For successful FMN-MS2 switches, the position of the FMN aptamer and MS2 hairpin were sequestered, and other nucleotides were embedded into stems. However, they demonstrate that across both the FMN-MS2 switches and the fluorescent aptamer switches, more diverse motif orderings and toggling of mechanisms results in optimal performance. If the position, motif ordering, and structure-toggling of RNA switch mechanisms strongly impact switch performance, the question arises as to how motif ordering might impact the design of riboswitches and riboregulators and whether other, more optimal architectures, have yet to be discovered.

Overall, the massive number of designs synthesized and validated through the workflow in Andreasson et al. (1) is a testament to the remarkable technologies for parallel synthesis of thousands of synthetic nucleic acid templates (12, 13), which can be transcribed both *in vitro* and *in vivo* into RNA. When coupled with screening technologies (14, 15), and deep-sequencing platforms with turn-around times in less than a day, the gap between *in silico* design and experimental validation will become smaller. Ultimately, the process will be limited only by the speed of DNA synthesis and the creativity in which we can think of new RNA switch structures.

The explosion of RNA data in recent years, with hundreds of thousands of RNA switches synthesized, validated, and released to the public, will allow for development of more quantitative and predictive theories for RNA structure design and eventually *de* novo, forward-engineered RNA functional design (16). The cycle of RNA design and testing can be short, especially compared to protein engineering. With the application of more sophisticated algorithms and motif finding in RNA regulatory elements, machine learning applications for RNA design will result in minimal experimental validation. Predicting the three-dimensional structure that a protein will adopt has been an important research problem for 50 years, but major advances have recently been achieved through artificial intelligence (AI) networks (17, 18). Only a few deep learning AI approaches have been developed for the computational prediction of three-dimensional RNA structures (19), but with the expansion and massively parallel high-throughput characterization of these datasets, AI predictions for RNA devices like riboswitches are close at hand.

In the age of big data, and with the growth of citizen science applications, the emergence of projects that are game-based and provide massive amounts of quantitative data make exploration of the complex sequence space of riboswitches and riboregulators feasible. Over time, machine learning or AI and the

development of more effective RNA design algorithms may mean that some types of data analysis may no longer require any human input. However, science-based games like the Foldit game (20), which rely on diversity in spatial recognition and problem-solving, should continue to provide valuable insights. As it currently stands, development of citizen science initiatives like Eterna, Foldit, or EyeWire (21), have enabled thousands of interested people to become involved in authentic scientific research from anywhere with internet connectivity.

Despite the ease with which games might be accessed and advancements in mobile technology, studies have shown that typical participants in citizen science initiatives are well-educated males with an existing interest in science or computing and are also likely to live in North America or Europe (23). Perhaps greater efforts targeting marginalized and minoritized audiences, groups who may feel that these opportunities aren't meant for them, may result in both greater number of participants, and may bring new perspectives and approaches to the data. While Andreasson et al. (1) make strides in "closing the loop" on crowd-sourced science to enable iterative hypothesis generation and experimental testing, by actively engaging and promoting Eterna to new, diverse audiences, the research might be more impactful if provisions are made to encourage participants to analyze their own data and ask their own questions. Opening research in an inclusive way may have wider societal benefits such as increasing transparency of research, while simultaneously helping build science capital, thus bringing populations who would otherwise be removed from the scientific process to the forefront of research discovery.

To cast the widest net, games built like Eterna must compete with the likes of Candy Crush, Pokemon GO, and other popular video games. If we build it, they do not necessarily come. In communities where there is low science capital, using online citizen science projects may help introduce scientific research to a range of ages and abilities, but the "democratization" of experimental biological science is far from a reality. The provisioning of a low-cost experimental validation workflow for hypotheses generated by the broader community is a crucial step toward making scientific discovery more accessible.

Acknowledgments

The author's research is supported by an NIH Director's New Innovator Award (DP2GM126892), NIH grants (U01AI148319, R01EB031893), an RCSA award (28422), and an NSF RAPID award (022329A).

Competing Interests

A.A.G. is a cofounder of En Carta Diagnostics Inc.

References

1. J. O. L. Andreasson, *et al.*, "Crowdsourced RNA design discovers diverse, reversible, efficient, self-contained molecular sensors" *Proc. Natl. Acad. Sci. U.S.A.* unknown in press. (2022)

2. M. N. Win, C. D. Smolke, Higher-Order Cellular Information Processing with Synthetic RNA Devices. *Science* **322**, 456–460 (2008).

3. S. Topp, *et al.*, Synthetic Riboswitches That Induce Gene Expression in Diverse Bacterial Species. *Applied and Environmental Microbiology* **76**, 7881–7884 (2010).

4. B. Townshend, A. B. Kennedy, J. S. Xiang, C. D. Smolke, High-throughput cellular RNA device engineering. *Nat Methods* **12**, 989–994 (2015).

5. K. Pardee, *et al.*, Rapid, Low-Cost Detection of Zika Virus Using Programmable Biomolecular Components. *Cell* **165**, 1255–1266 (2016).

6. A. A. Green, *et al.*, Complex cellular logic computation using ribocomputing devices. *Nature* **548**, 117–121 (2017).

7. J. N. Zadeh, *et al.*, NUPACK: Analysis and design of nucleic acid systems. *Journal of Computational Chemistry* **32**, 170–173 (2011).

8. R. Lorenz, *et al.*, ViennaRNA Package 2.0. *Algorithms for Molecular Biology* **6**, 26 (2011).

9. A. A. Green, P. A. Silver, J. J. Collins, P. Yin, Toehold Switches: De-Novo-Designed Regulators of Gene Expression. *Cell* **159**, 925–939 (2014).

10. J. Lee, *et al.*, RNA design rules from a massive open laboratory. *Proceedings of the National Academy of Sciences* **111**, 2122–2127 (2014).

11. M. J. Wu, J. O. L. Andreasson, W. Kladwang, W. Greenleaf, R. Das, Automated Design of Diverse Stand-Alone Riboswitches. *ACS Synth. Biol.* **8**, 1838–1846 (2019).

12. X. Zhou, *et al.*, Microfluidic PicoArray synthesis of oligodeoxynucleotides and simultaneous assembling of multiple DNA sequences. *Nucleic Acids Res* **32**, 5409–5417 (2004).

13. E. M. LeProust, *et al.*, Synthesis of high-quality libraries of long (150mer) oligonucleotides by a novel depurination controlled process. *Nucleic Acids Research* **38**, 2522–2540 (2010).

14. J. B. Lucks, *et al.*, Multiplexed RNA structure characterization with selective 2′-hydroxyl acylation analyzed by primer extension sequencing (SHAPE-Seq). *Proceedings of the National Academy of Sciences* **108**, 11063–11068 (2011).

15. J. N. Pitt, A. R. Ferré-D'Amaré, Rapid Construction of Empirical RNA Fitness Landscapes. *Science* **330**, 376–379 (2010).

16. N. M. Angenent-Mari, A. S. Garruss, L. R. Soenksen, G. Church, J. J. Collins, A deep learning approach to programmable RNA switches. *Nat Commun* **11**, 5057 (2020).

17. J. Jumper, *et al.*, Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).

18. M. Baek, *et al.*, Accurate prediction of protein structures and interactions using a three-track neural network. *Science* **373**, 871–876 (2021).

19. J. A. Valeri, *et al.*, Sequence-to-function deep learning frameworks for engineered riboregulators. *Nat Commun* **11**, 5058 (2020).

20. S. Cooper, *et al.*, Predicting protein structures with a multiplayer online game. *Nature* **466**, 756–760 (2010).

21. J. S. Kim, *et al.*, Space–time wiring specificity supports direction selectivity in the retina. *Nature* **509**, 331–336 (2014).

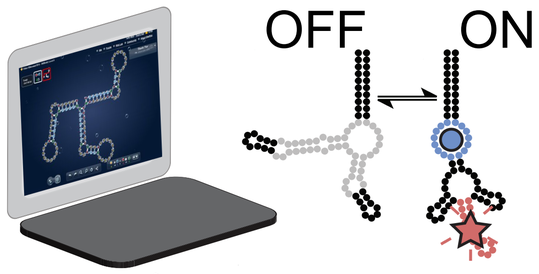
22. V. Curtis, *Online Citizen Science and the Widening of Academia* (Springer International Publishing, 2018) https:/doi.org/10.1007/978-3-319-77664-4 (April 25, 2022).

23. V. Curtis, "Who Takes Part in Online Citizen Science?" in *Online Citizen Science and the Widening of Academia: Distributed Engagement with Research and Knowledge Production*, Palgrave Studies in Alternative Education., V. Curtis, Ed. (Springer International Publishing, 2018), pp. 45–68.

FIGURE CAPTIONS

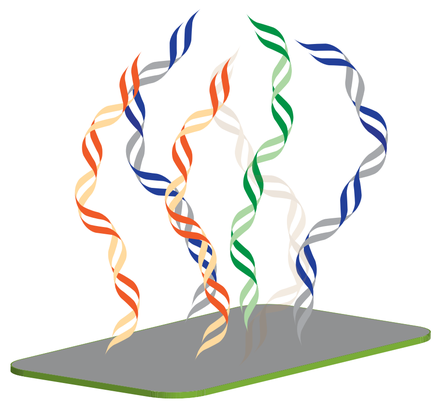
Fig 1. Iterative cycle of community-based hypothesis testing. Videogame-based crowdsourcing of RNA designs through Eterna is followed by synthesis of RNA libraries tested in massively parallel fashion on an array. High-throughput functional characterization and data quantification is released back to the Eterna community, allowing for iterative rounds of hypothesis generation, modification, and analysis by citizen scientists.

Eterna



Community Hypothesis Generation

High-Throughput RNA Synthesis



GCCCUUAUUC...
CCACCCACCC...
CUCCCUUUCC...
CCCACCUCCA...

RNA Design Testing