Published online 14 June 2022 *Nucleic Acids Research, 2022, Vol. 50, Web Server issue* ***W639–W650***
https://doi.org/10.1093/nar/gkac508

REVERSE: a user-friendly web server for analyzing next-generation sequencing data from *in vitro* selection/evolution experiments

Zoe Weiss[1] and Saurja DasGupta[2,3,4,*]

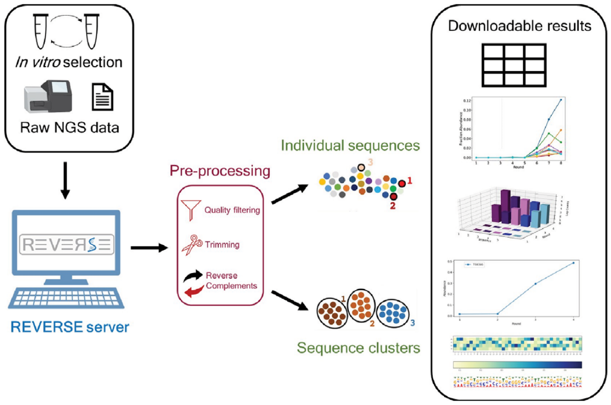
[1]Department of Chemistry and Chemical Biology, Harvard University, Cambridge, MA 02138, USA, [2]Department of Molecular Biology, Center for Computational and Integrative Biology, Massachusetts General Hospital, Boston, MA 02114, USA, [3]Howard Hughes Medical Institute, Massachusetts General Hospital, Boston, MA 02114, USA and [4]Department of Genetics, Harvard Medical School, Boston, MA 02115, USA

Received March 21, 2022; Revised May 22, 2022; Editorial Decision May 27, 2022; Accepted May 27, 2022

ABSTRACT

Next-generation sequencing (NGS) enables the identification of functional nucleic acid sequences from *in vitro* selection/evolution experiments and illuminates the evolutionary process at single-nucleotide resolution. However, analyzing the vast output from NGS can be daunting, especially with limited programming skills. We developed REVERSE (Rapid EValuation of Experimental RNA Selection/Evolution) (https://www.reverseserver.org/), a web server that implements an integrated computational pipeline through a graphical user interface, which performs both pre-processing and detailed sequence level analyses within minutes. Raw FASTQ files are quality-filtered, dereplicated, and trimmed before being analyzed by either of two pipelines. The first pipeline counts, sorts, and tracks enrichment of unique sequences and user-defined sequence motifs. It also identifies mutational intermediates present in the sequence data that connect two input sequences. The second pipeline sorts similar sequences into clusters and tracks enrichment of peak sequences. It also performs nucleotide conservation analysis on the cluster of choice and generates a consensus sequence. Both pipelines generate downloadable spreadsheets and high-resolution figures. Collectively, REVERSE is a one-stop solution for the rapid analysis of NGS data obtained from *in vitro* selection/evolution experiments that obviates the need for computational expertise.

GRAPHICAL ABSTRACT



INTRODUCTION

In vitro selection/evolution is a powerful technique to isolate nucleic acids with desired functions without a priori knowledge of their sequences by imposing selection pressure upon combinatorial libraries (1–3). Functional nucleic acids such as aptamers and (deoxy)ribozymes identified by this technique have been used in diagnostics and therapy (4–6), and *in vitro* evolution has been used to study the intricacies of molecular evolution (7–14) and explore the biochemical capabilities of nucleic acids in the context of the origin of life (15–17). With the advent of next-generation sequencing (NGS), it has been possible to investigate the outcomes of combinatorial selections in unprecedented detail as NGS provides expansive sequence coverage. In addition to identifying hundreds of thousands of sequences that possess the desired phenotype, analysis of NGS data from each round may reveal the evolutionary trajectory of each sequence under changing selection pressures, thereby generating a high-resolution picture of the evolutionary process.

Sequence files obtained from NGS are too large to be manipulated without the use of the command-line or spe-

[*]To whom correspondence should be addressed. Tel: +1 617 726 5102; Fax: +1 617 643 3328; Email: dasgupta@molbio.mgh.harvard.edu

cialized software. As a result, it is practically impossible to extract the most basic, yet essential information needed to biochemically validate selection outcomes without considerable bioinformatics expertise. NGS data must first be pre-processed to filter out 'low quality' sequences and remove sequencing adapters and constant regions (such as primer binding sites). Next, identical sequences must be binned (dereplication) in order to identify and count all the different sequences present in the data (unique sequences) and converted to their reverse complements, if analyzing reverse reads. Usually, an arbitrary subset of the most abundant sequences is chosen for biochemical characterization. However, in order to sample the diversity of selected sequences, it is beneficial to cluster closely related sequences and characterize the most abundant sequence from each cluster (referred to as peak sequence). NGS data from all rounds of selection can be used to quantify sequence enrichment, investigate evolutionary fitness landscapes and identify nucleotides important for function through nucleotide conservation analysis. However, these results are inaccessible to experimentalists without significant programming experience. Therefore, the field has largely depended on custom computational pipelines created by individual labs through collaborations with bioinformaticians. The unavailability of a standardized pipeline to analyze NGS data from *in vitro* selection/evolution experiments and more importantly, the absolute reliance on programming expertise that most experimentalists lack, have limited the wide implementation of NGS technology to combinatorial selections.

We developed REVERSE (Rapid EValuation of Experimental RNA Selection/Evolution) with the goal of democratizing bioinformatic analysis of NGS data from *in vitro* selection/evolution experiments by removing all computational barriers. REVERSE is an integrated computational pipeline that runs a set of custom scripts written in Python and interacts with the user through an intuitive web-based graphical interface. The entire pipeline requires a single upload step at the beginning where the user uploads a raw FASTQ file for each selection round. No subsequent upload is necessary for downstream tasks. The pre-processing module generates spreadsheets containing quality-filtered, trimmed, and dereplicated sequences with their read counts. Then the user is offered two pipelines—the first analyzes individual sequences and the second clusters similar sequences before analyzing sequence clusters. REVERSE identifies the most abundant sequences/peak sequences of the most abundant clusters and quantifies their enrichment during selection. More advanced functionalities like identification of mutational intermediates between two sequences, sequence motif search, or conservation analysis are also available. In addition to spreadsheets, REVERSE outputs downloadable publication-quality figures for most results. To demonstrate its utility, we applied the REVERSE workflow to a subset of NGS data (four rounds) obtained from a previous *in vitro* selection experiment designed to isolate ligase ribozymes (18). The entire REVERSE pipeline to analyze 100 000 sequences per round took about 22 minutes to complete, which includes about 4 minutes for file upload and pre-processing. In addition to analyzing outcomes from nucleic acid selection/evolution, REVERSE may be used for the computational needs of other experiments that involve combinatorial libraries such as peptide and protein selection using mRNA (19) or phage display (20) and high-throughput mutational analysis of (deoxy)ribozymes (21). Therefore, REVERSE represents the first dedicated web-based computational platform for rapid analysis of NGS data acquired from combinatorial selections that enables users to expand the use of their experimental results without having to write a single line of code.

MATERIALS AND METHODS

Overview

The *Analyze* tab is the gateway to the REVERSE workflow where sequence data files are pre-processed and analyzed in two separate pipelines – the first pipeline analyzes individual sequences and the second pipeline clusters sequences before analyzing individual clusters (Figure 1). REVERSE takes as input one zipped FASTQ file (e.g. 'filename.fastq.zip') per round and does not merge paired-end reads from forward and reverse runs. As most combinatorial libraries are <100 nt, single-end reads are usually sufficient to cover the library sequence. REVERSE offers two types of outputs—spreadsheets (CSV files) and downloadable visualizations. Spreadsheets contain lists of sequences with their respective read counts and selection statistics. Spreadsheets allow easy manipulation of sequence data and importantly, enable instant identification of selected sequences for further biochemical validation. Visualizations include high-resolution line plots, 3D bar plots, heatmaps, and sequence logos. All visualizations are accompanied by raw data that may be used to create customized figures. Tables can be downloaded as CSV files with a single click and figures can be downloaded as PNG files from the web page by right-clicking on the image and selecting the 'Save Image As' option.

Pre-processing module: Filtering by quality, dereplication, trimming and batch conversion to reverse complements, if required

The first page of the *Analyze* tab contains the first four steps of the pre-processing module that determine the operations to be performed on the sequence data files, once uploaded (Figure 2). *Step 1* allows the user to input the number of files they want analyzed (usually corresponding to the number of rounds). REVERSE can analyze a maximum of ten files at once; however, computational time is reduced with fewer files. For experiments with more than ten rounds, users may analyze data from representative rounds that capture functional enrichment or analyze every other round, if possible. Sequencing results from replicate experiments or from parallel selections performed under different selection conditions can also be analyzed via the multiple file upload functionality of REVERSE. *Step 2* reads the base calling accuracy at each nucleotide position (Q score) of every sequence and filters out sequences that have base calling errors larger than the user-defined cutoff. For example, a cutoff input of 98% will filter out sequences that have > 2% of their nucleotides with a *Q* score <20 (i.e. >1% error in base calling). *Step 3* accepts the number of sequences in each file the user wants analyzed, which could be the first

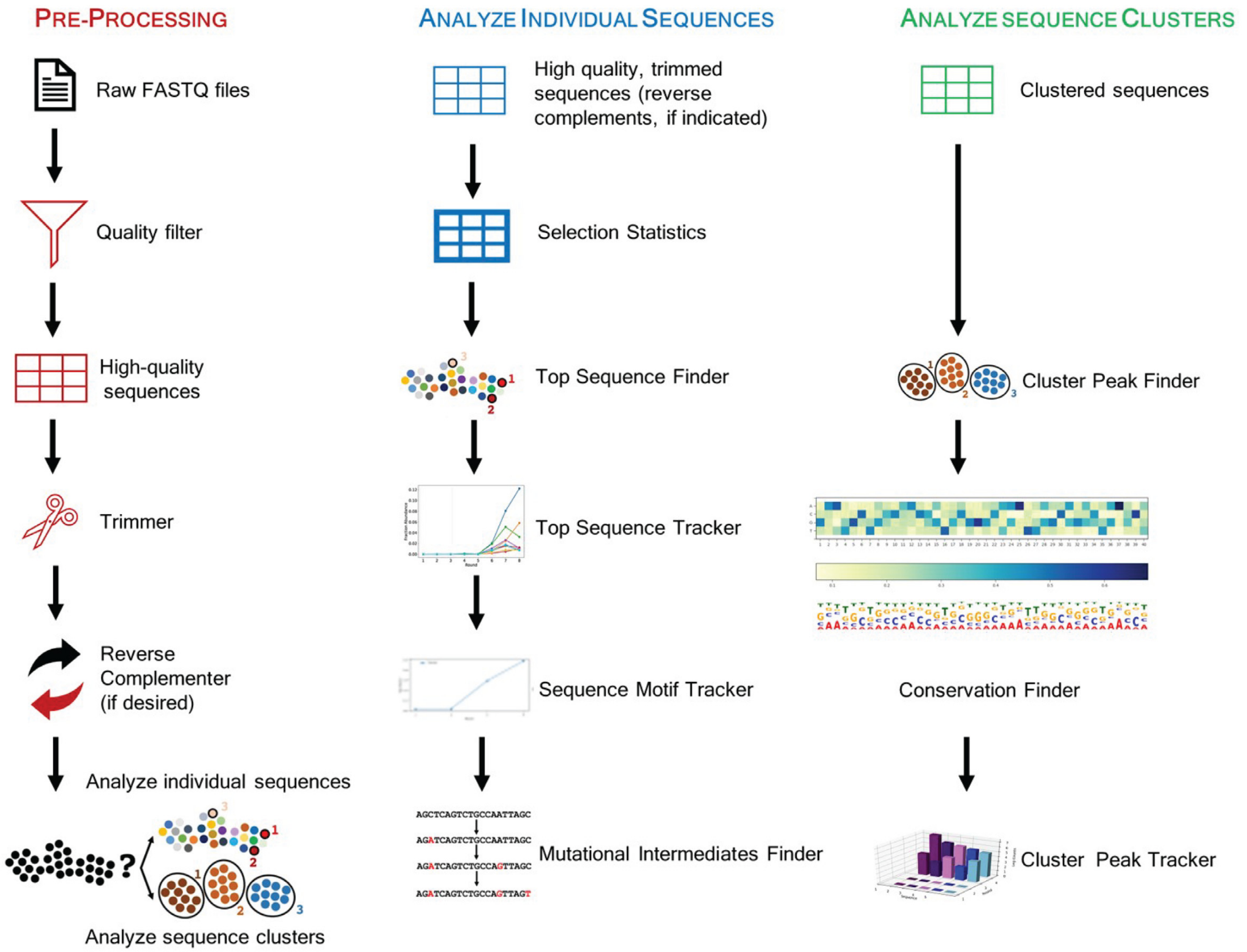
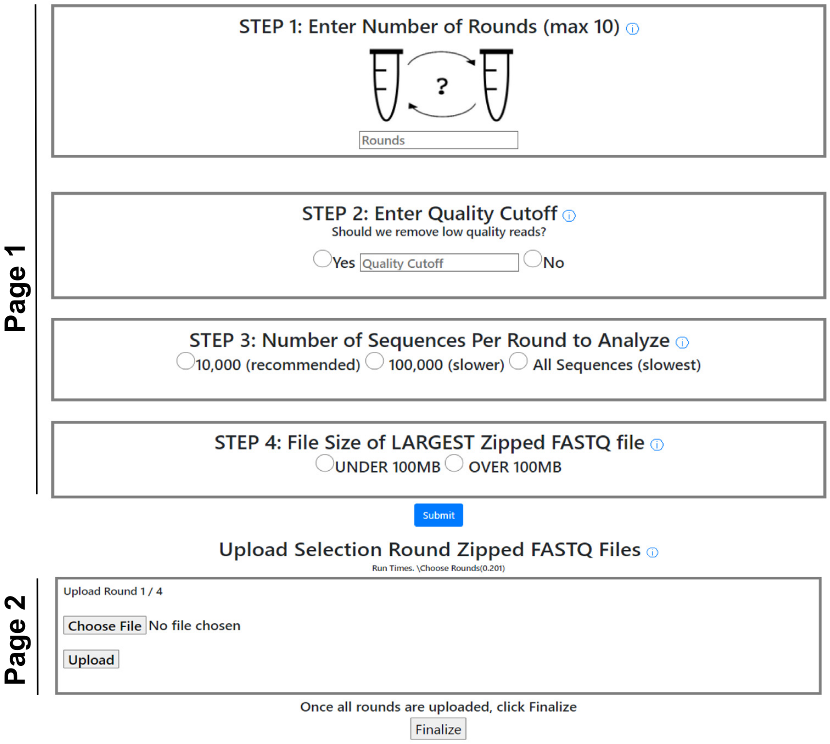


Figure 1. The REVERSE workflow: from raw next-generation sequencing data to data visualization with a few clicks. REVERSE pre-processes raw FASTQ files from combinatorial selections and rapidly identifies the most abundant sequences. It clusters similar sequences and identifies the most abundant clusters, and tracks the enrichment of individual sequences, cluster peak sequences, and user-defined sequence motifs. REVERSE also performs conservation analysis on sequence clusters and searches sequence data for intermediates between two user-defined sequences.

10 000, the first 100 000, or all sequences. Analyzing a subset of the sequence data lowers processing time while preserving important information such as the identities of the most abundant sequences/sequence clusters and their relative abundances and enrichment during selection (see Results). In *Step 4*, the user enters the size of the input ZIP files to determine the upload strategy. REVERSE can directly accept compressed sequence files under 100MB; however, files larger than 100MB can be accommodated in REVERSE with an extra processing step. The second page of the *Analyze* tab permits the user to upload multiple FASTQ files to be analyzed simultaneously. Files under 100MB can be directly uploaded here (Figure 2). For files over 100MB, both Mac and Windows users may use the Galaxy server to filter out low quality sequences first and use a compressed form of this smaller output file as input for REVERSE. Alternatively, Mac users may download a simple bash script to split large files (PC users may need to use a terminal emulator). This script splits and downloads these files to the user's desktop. For example, this script will generate two smaller files, 'roundXaa' and 'roundXab' from 'roundX', a single >100MB file. The user can then upload these smaller files in the *Upload Split Zipped FASTQ Files* page, where REVERSE will merge the two files. All the commands needed for this operation are provided on this page if the user chooses the 'over 100MB' option in Step 4. As the script currently splits a single file into two smaller files and the size limit for each file upload is 100MB, the size limit for the larger file is 200MB, although this can be increased by updating the script so that a large file is automatically split into the required number of smaller files, each under 100MB.

Once all files are uploaded, REVERSE removes sequences with quality scores below the user-defined cutoff, bins identical sequences, and counts the number of times each unique sequence was read for each file. *Step 5a* gives the user the option to view the 'high-quality' sequences for each uploaded sequence file in downloadable CSV files (Figure 3). This is particularly helpful in identifying the region of interest in each sequence so that flanking constant regions such as barcode sequences and primer binding sequences can be removed. *Step 5b* allows the user to define



STEP 1: Enter Number of Rounds (max 10)

Rounds

STEP 2: Enter Quality Cutoff

Should we remove low quality reads?

○Yes Quality Cutoff ○No

STEP 3: Number of Sequences Per Round to Analyze

○10,000 (recommended) ○ 100,000 (slower) ○ All Sequences (slowest)

STEP 4: File Size of LARGEST Zipped FASTQ file

○UNDER 100MB ○ OVER 100MB

Submit

Upload Selection Round Zipped FASTQ Files

Run Times. \Choose Rounds(0.201)

Upload Round 1 / 4

Choose File No file chosen

Upload

Once all rounds are uploaded, click Finalize

Finalize

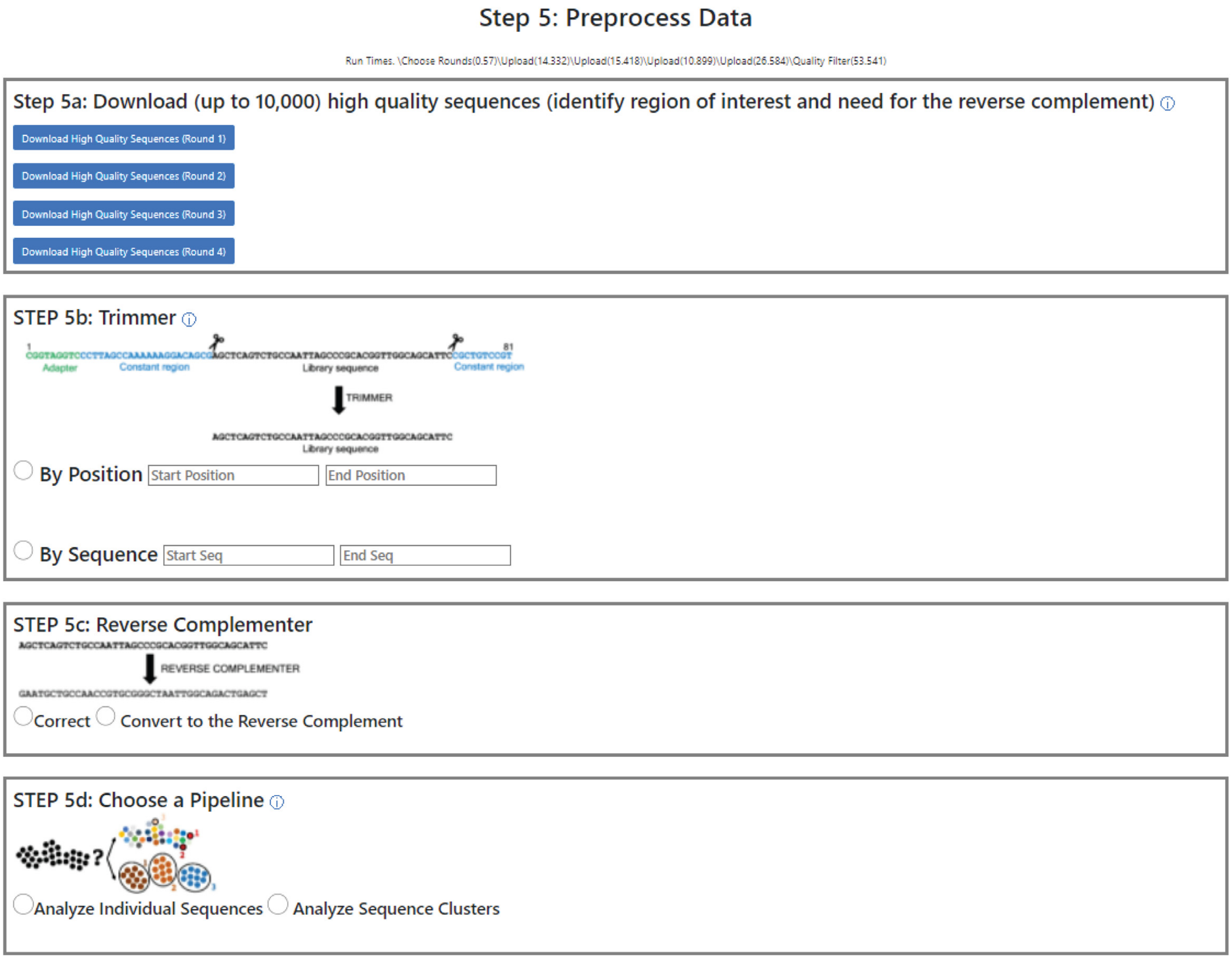
Figure 2. The pre-processing module of REVERSE: Input. Page 1 sets up the parameters for subsequent analysis and page 2 provides the uploading interface for compressed FASTQ files. The pre-processing module bins identical sequences (dereplication) and calculates the number of reads for each dereplicated sequence.

the boundary of the region of interest by either entering its first and last nucleotide numbers or the flanking constant primer/adapter sequences in the *Trimmer* tool. In *Step 5c,* the user indicates if the uploaded files are from forward or reverse sequencing runs. The *Reverse Complementer* tool performs batch conversion of all sequences derived from reverse sequencing runs. Finally, in *Step 5d*, the user selects one of two pipelines of the analysis module of REVERSE. While the *Analyze Individual Sequences* pipeline performs a series of computations on all dereplicated sequences, the *Analyze Sequence Clusters* pipeline first clusters closely related dereplicated sequences before analyzing the most abundant sequence clusters.

Analysis module: analyze individual sequences

The *Analyze Individual Sequences* pipeline consists of the following six tools (Figure 4):

1. *Download Pre-Processed Data* outputs a CSV file, for each input FASTQ file, containing quality-filtered sequences that are trimmed to the desired length, and, if desired, converted to their reverse complements. Each output file contains up to 10 000 dereplicated sequences that are sorted according to their read counts. This tool provides rapid access to sequences isolated from selection and enables the user to inspect and manipulate sequencing data in a spreadsheet. This is one of the most sought-after results in most combinatorial selections.
2. *Selection Statistics* outputs a table containing the number of total sequences, number of high-quality sequences, number of unique sequences, and percent of unique sequences for each round. This tool may be used to evaluate selection outcomes. A decrease in the fraction of unique sequences indicates sequence enrichment, which is usually accompanied by functional enrichment and could point to successful selection.
3. *Top Sequence Finder* outputs the *N* most abundant sequences from the final round of selection with their read counts, where *N* is user-defined. This tool allows the identification of the subset of abundant sequences the user plans to characterize biochemically. These sequences are often the most active (i.e. possess the highest fitness) for the desired function.
4. *Top Sequence Tracker* plots the fractional abundance of each of the *N* most abundant sequences against selection round. In addition to visualizing sequence-level population dynamics of selection/evolution across rounds, this tool visually depicts the point in the selection where enrichment becomes evident. The tool can therefore be used to monitor selection progress even before the detection of the desired activity in selection pools. Since this tool compares the fractional abundance of the most abundant sequences across sequence files, it can be used to test selection reproducibility (where files represent experimental replicates instead of outputs from different rounds) or analyze the differences in sequence enrichment under diverse selection pressures (where each file represents a different selection condition).



Step 5: Preprocess Data

Run Times. \Choose Rounds(0.57)\Upload(14.332)\Upload(15.418)\Upload(10.899)\Upload(26.584)\Quality Filter(53.541)

Step 5a: Download (up to 10,000) high quality sequences (identify region of interest and need for the reverse complement) ⓘ

Download High Quality Sequences (Round 1)
Download High Quality Sequences (Round 2)
Download High Quality Sequences (Round 3)
Download High Quality Sequences (Round 4)

STEP 5b: Trimmer ⓘ

1
81
CGGTAGGTCCCTTAGCCAAAAAAGGACAGCGAGCTCAGTCTGCCAATTAGCCCGCACGGTTGGCAGCATTCCGCTGTCCGT
Adapter
Constant region
Library sequence
Constant region
TRIMMER
AGCTCAGTCTGCCAATTAGCCCGCACGGTTGGCAGCATTC
Library sequence

○ By Position Start Position End Position

○ By Sequence Start Seq End Seq

STEP 5c: Reverse Complementer

AGCTCAGTCTGCCAATTAGCCCGCACGGTTGGCAGCATTC
REVERSE COMPLEMENTER
GAATGCTGCCAACCGTGCGGGCTAATTGGCAGACTGAGCT

○ Correct ○ Convert to the Reverse Complement

STEP 5d: Choose a Pipeline ⓘ

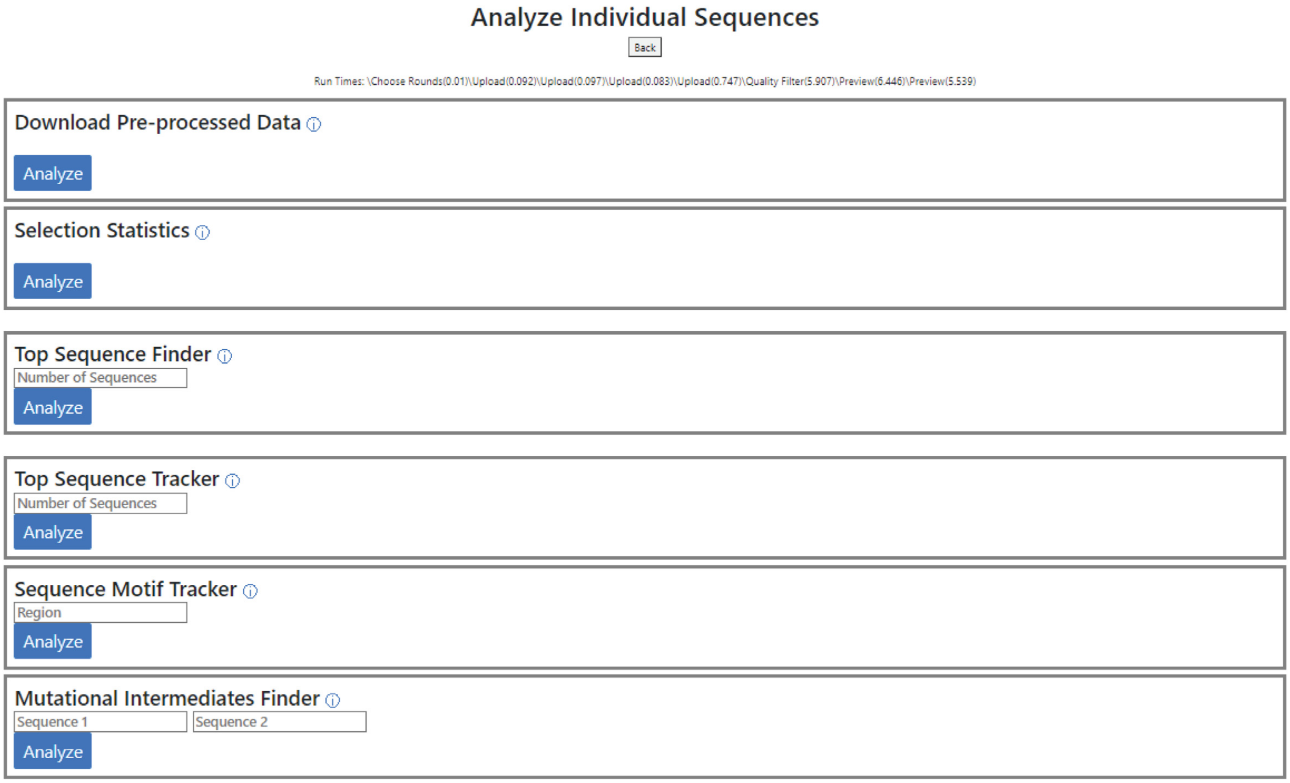
○ Analyze Individual Sequences ○ Analyze Sequence Clusters

Figure 3. The pre-processing module of REVERSE: output. Step 5a provides a list of dereplicated sequences, the Trimmer tool in Step 5b removes undesired regions from these sequences, and the Reverse Complementer tool in Step 5c converts sequences to their reverse complements if the user desires. This tool is useful for analyzing reverse reads. Finally, Step 5d offers the user two pipelines of the analysis module of REVERSE.

5. *Sequence Motif Tracker* outputs a spreadsheet for each round containing only sequences that possess the user-defined motif and plots the fractional abundances of these sequences across rounds. A sequence motif in this context is defined as a continuous string of nucleotides and is distinct from a structural motif such as a pseudoknot or a stem–loop, which can be formed by divergent sequences. Sequence motifs must be defined using standard nomenclature – specific nucleotides: A, T, G, C; purines: R, pyrimidines: Y, and any nucleotide: N. Functional enrichment during selection is often accompanied by the enrichment of sequences that possess specific sequence motifs. These sequence motifs may constitute structural features important for activity and/or directly interact with the substrate (in case of (deoxy)ribozyme selections).
6. *Mutational Intermediates Finder* takes as input two sequences and performs an advanced search across all sequence files to identify sequence variants that lie in the sequence space between the two. These sequence variants may differ from each other by one or more point mutations depending on the extent of sequence space coverage in the specific experiment. The intermediates identified by this tool must be present in the sequencing data files uploaded to REVERSE, therefore, neutral pathways between two input sequences can only be identified if all the required intermediate sequences are present in the uploaded sequence data. With experiments that enable complete or near-complete coverage of sequence space (11), *Mutational Intermediates Finder* can potentially identify neutral networks between distinct functional RNAs. The existence of neutral networks between functional RNAs such as ribozymes has significant implications for the evolution of RNA-based early life (22–24).

Analysis module: analyze sequence clusters

Closely related RNA sequences usually adopt similar structures and therefore can be considered to be members of the same functional RNA class. To identify multiple classes of functional RNAs for downstream characterization, se-



Analyze Individual Sequences

Run Times: \Choose Rounds(0.01)\Upload(0.092)\Upload(0.097)\Upload(0.083)\Upload(0.747)\Quality Filter(5.907)\Preview(6.446)\Preview(5.539)

Download Pre-processed Data ⓘ

Selection Statistics ⓘ

Top Sequence Finder ⓘ

Top Sequence Tracker ⓘ

Sequence Motif Tracker ⓘ

Mutational Intermediates Finder ⓘ

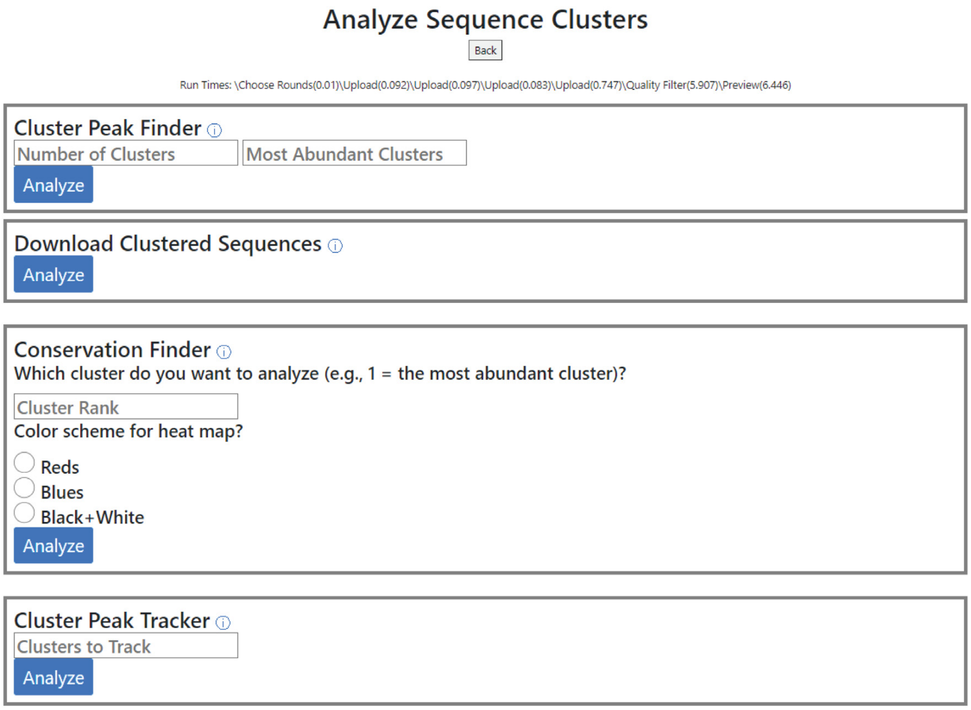
Figure 4. The *Analyze Individual Sequences* pipeline of the analysis module of REVERSE. This pipeline allows the user to download pre-processed data in easy-to-use spreadsheets that contain a list of unique sequences sorted according to their abundances, accompanied by their read counts. The Selection Statistics tool generates a table summarizing the selection outcome. The Top Sequence Finder tool identifies the *N* (user-defined) most abundant sequences in the last round data and the Top Sequence Tracker tool tracks the enrichment of the abundant sequences across rounds. The Sequence Motif Tracker tool searches all uploaded sequence files for a user-defined nucleotide string and outputs a spreadsheet for each round that contains only sequences with the desired nucleotide string. It also generates a graphic illustrating the enrichment of sequences containing this motif across rounds.

lected sequences are often clustered through multiple sequence alignments and the most abundant sequence (peak sequence) from each of the most abundant clusters, is characterized. To address this need, we created the *Analyze Sequence Clusters* pipeline which consists of the following four tools (Figure 5):

1. *Cluster Peak Finder* outputs peak sequences (with their read counts) from *K* sequence clusters, where *K* is user-defined ($K \leq 50$). The most abundant peak sequences usually represent distinct solutions for accomplishing the selected function. This tool, therefore, provides easy access to the most abundant clusters and their peak sequences, which is one of most desired results from *in vitro* selection experiments. While clustering accuracy increases with increase in the number of clusters, it is more time consuming. As this tool, like other clustering and multiple sequence alignment software, is time and memory intensive, currently it is unable to cluster sequences from several large files at once. Therefore, if using large files, the user may perform clustering with one or two input files at once. This is not a problem for most users who would likely analyze data from only the final round of selection.
2. *Download Clustered Sequences* allows the user to access sequence clusters. This tool outputs CSV files containing up to 1000 sequences (with read counts) that are assigned to specific clusters (cluster 0 onwards). Using the Sort function in applications like Microsoft Excel, the user can organize sequences according to their cluster number and further sort sequences in a cluster according to their read counts.
3. *Conservation Finder* outputs a customizable heatmap that illustrates the distribution of nucleotides at each position within a user-specified cluster (*R)* and a sequence logo for the consensus sequence derived from this data. Here *R* denotes the cluster rank, i.e. 1 denotes most abundant cluster, 2, the second most abundant cluster as so on. Nucleotide variations at each position may be used to predict the importance of each nucleotide to the sequence's fitness.
4. *Cluster Peak Tracker* outputs a plot containing the fractional abundances of the peak sequences of the *N* most abundant clusters across rounds. Like *Top Sequence Tracker*, this tool can be used to evaluate selection progress before desired activity is detected in sequence pools and obtain information about the sequence population dynamics.

Implementation

The REVERSE web server is built using PythonAnywhere, an online integrated development platform running server-based Python scripts for web development. The back end of



Analyze Sequence Clusters

Back

Run Times: \Choose Rounds(0.01)\Upload(0.092)\Upload(0.097)\Upload(0.083)\Upload(0.747)\Quality Filter(5.907)\Preview(6.446)

Cluster Peak Finder ⓘ

Number of Clusters | Most Abundant Clusters

Analyze

Download Clustered Sequences ⓘ

Analyze

Conservation Finder ⓘ

Which cluster do you want to analyze (e.g., 1 = the most abundant cluster)?

Cluster Rank

Color scheme for heat map?

○ Reds
○ Blues
○ Black+White

Analyze

Cluster Peak Tracker ⓘ

Clusters to Track

Analyze

Figure 5. The *Analyze Sequence Cluster* pipeline of the analysis module of REVERSE. This pipeline allows the user to download a spreadsheet containing pre-processed sequence data with closely related sequences sorted into clusters. Each sequence is assigned a cluster number. Cluster Peak Finder outputs the peak sequences (and their read counts) from the *N* (user-defined) most abundant clusters. Conservation Finder quantifies the distribution of each of the four nucleotides at each position within the cluster of choice and outputs the results in a customizable heatmap and associated sequence logo. Cluster Peak Tracker quantifies the enrichment of peak sequences of the *N* most abundant clusters.

REVERSE is created in Python (https://www.python.org/downloads/release/python-3101/) with Flask (https://flask.palletsprojects.com/en/2.0.x/). The various computational tools in REVERSE are written as custom scripts in Python. Comprehensive documentation can be found in the Supplementary Data file and on the web page, which also includes a video walkthrough with demo files. The front end of REVERSE consists of a graphical user interface created with HTML and CSS scripts in PythonAnywhere. The basic layout of each page is based on designs from the Bootstrap v5.1 library (https://getbootstrap.com/). All data visualizations in REVERSE are generated using the Matplotlib and Seaborn libraries in Python. REVERSE has been tested with various combinations of Linux/ macOS/Windows 10 and Chrome/Firefox/Edge/Safari (Supplementary Table S1). Release notes containing updates to the various tools of REVERSE are provided on the website and associated Github and will continue to be updated.

Comparison with similar software

REVERSE is the only web server that performs both pre-processing and dedicated sequence analysis of NGS data obtained from combinatorial selections. Although software/web servers catering to RNA-seq data analysis for biological applications such as differential gene expression exist (25,26), there are next to no resources for analyzing sequence files from *in vitro* selection/evolution experiments, and therefore multiple software must be used. For example, a collection of command-line software is often used to pre-process raw FASTQ files. Quality filtering may be performed using FastQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/), constant regions may be removed using Cutadapt, cutPrimers, or Trimmomatic (27–29) and paired-end reads may be merged using PANDAseq (30) or PEAR (31). The Galaxy server provides a web-based GUI to implement some of these tools. Advanced users may create custom analytical workflows by catenating suitable tools; however, this requires familiarity with the many bioinformatic tools that are available in Galaxy. A direct workflow that caters to the needs of combinatorial selection is not obvious to the user. In fact, the versatility of the Galaxy server in analyzing sequence data from diverse experiments can be overwhelming to the user who wants to obtain very specific results that are unique to combinatorial selections. REVERSE simplifies the analytical process by presenting its bioinformatic algorithms as interactive tools. Its self-contained workflow allows the user to perform analyses specifically geared toward selection experiments without the distraction of options that are not relevant to their analyses.

EasyDIVER (32), a command-line software, can perform paired-end joining, dereplication and trimming, but performs no further bioinformatic analysis. FASTAptamer (33), another command-line software, consisting of modular Perl scripts, was developed to specifically analyze NGS data from combinatorial selections. It counts dereplicated sequences, compares relative abundances in sequences present in two/three files, searches for degenerate sequence motifs, and performs sequence clustering. However, FASTAptamer does not contain any pre-processing functionalities, and its command-line interface limits wide adoption by experimentalists not familiar with the command line. Further, FASTAptamer primarily outputs results within the command-line and does not generate easily usable sequence lists or visualizations. In addition to providing a more user-friendly interface (GUI versus command-line), REVERSE performs several computational tasks that the current version of FASTAptamer is unable to perform. These include the ability to analyze multiple sequence files simultaneously (FASTAptamer can do a maximum of three files), which allows the user to track sequence enrichment across selection. This enables the quantification of sequence level population dynamics and more importantly helps the user evaluate the trajectory of their experiment. Unlike FASTAptamer, REVERSE performs nucleotide conservational analysis on individual clusters, which provides information about the importance of each nucleotide in the nucleic acid's biochemical/biological function. To compare similar functionalities between the two software, we analyzed two sequence files provided by Alam *et al.* (in the manuscript reporting FASTAptamer) (33) using both REVERSE and FASTAptamer. REVERSE and FASTAptamer generated comparable results even when a subset of sequences (100 000 sequences) was analyzed by REVERSE. For example, the three most abundant sequences identified by FASTAptamer were present in the top five sequences identified by REVERSE and their enrichment (fraction abundance in file 1/fraction abundance in file 2) were similar. Similar functionalities are present within AptaSuite (34), which can be implemented as a locally-run GUI-based software, AptaGUI (35). AptaGUI additionally performs secondary structure prediction for individual sequences or sequence clusters to generate consensus structures. While FASTAptamer and AptaSuite perform similar functions as REVERSE, an intuitive and easy-to-use web server such as REVERSE has the potential to reach more users than any locally-run software. This is perhaps best illustrated in the area of nucleic acid secondary structure prediction by the mFOLD (now UNAfold) server.

REVERSE is a one-stop solution for the most common bioinformatic needs of combinatorial selections that interacts with users using a web-based GUI and requires no coding expertise or any familiarity with the command-line. By implementing custom scripts that run in the background, REVERSE allows users to upload multiple files in a single step at the beginning and freely use the tool of their choice without being constrained by the need to use files of specific formats at each step of analysis. In addition, REVERSE is specifically designed to readily output downloadable publication-quality visualizations by analyzing NGS data from combinatorial selections. Therefore, by simply uploading the raw FASTQ files obtained from sequencing and entering the required parameters in each step, users will be able to access the most important results from selection experiments, which will accelerate further experimental characterization and the provide publishable figures to accompany these results.

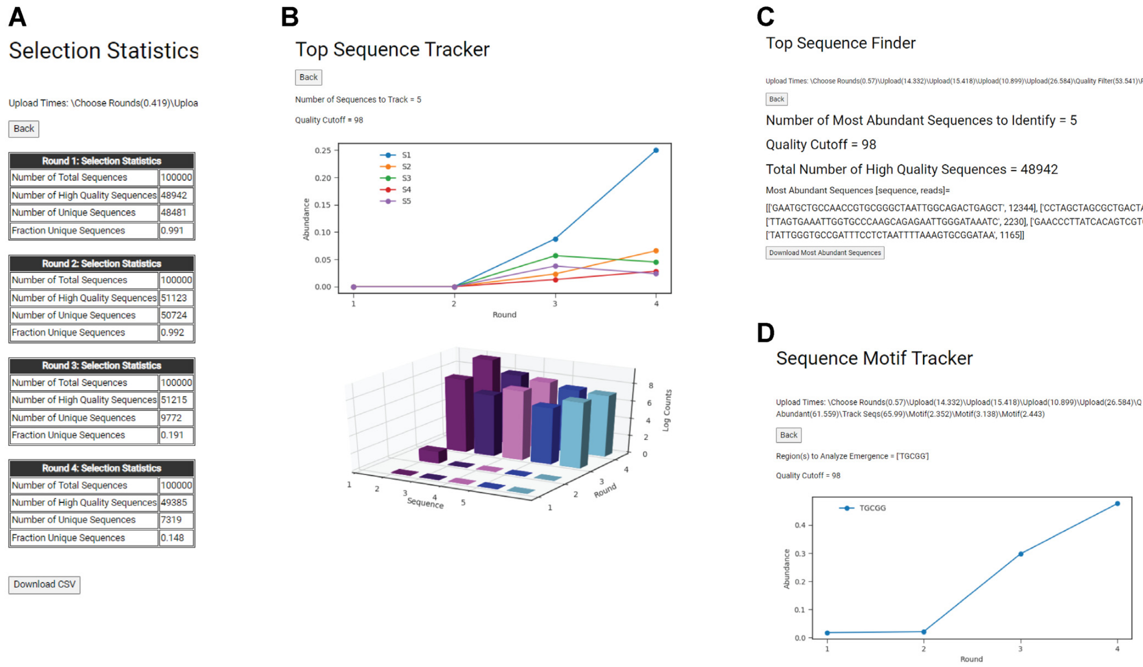
RESULTS

Case study: the reverse workflow

To illustrate the various functionalities of REVERSE, we analyzed four rounds (two before functional enrichment was detectable and two after) of raw NGS data from a previous *in vitro* selection experiment designed to isolate ribozymes that catalyze the templated ligation of RNA oligomers activated with 2-aminoimidazole (18). Users may try out all the tools REVERSE has to offer in the *Tutorial* tab with truncated versions of these files (demo files). All results discussed here have been obtained from REVERSE 1.2. After entering pre-processing parameters in *Steps 1–4* like number of rounds (4), quality cutoff (98), number of sequences to be analyzed for each file (100 000) and file size (under 100MB), four FASTQ files corresponding to reverse reads from each of the four rounds were uploaded to the pre-processing module. After inspecting high-quality sequences in downloaded spreadsheets (one for each round) in *Step 5a*, the region of interest (representing the 40 nt randomized library) was identified between nucleotides 32 and 71. This information was used to remove extraneous sequences in *Step 5b*. Since the uploaded sequence data was obtained from reverse runs, all sequences were converted to their reverse complements in *Step 5c*. In *Step 5d*, the *Analyze Individual Sequences* pipeline was selected first.

We downloaded pre-processed data for each round as separate spreadsheets, which allowed us to readily identify the most abundant sequences from each round. Selection outcomes from the *Selection Statistics* tool revealed a marked decrease in the fraction of unique sequences (from 0.991 in round 1 to 0.148 in round 4) indicating significant sequence enrichment (Figure 6A). The *Top Sequence Finder* and *Top Sequence Tracker* tools were used to output the five most abundant sequences in the round 4 sequence data and track its fractional abundance through rounds 1–4 (Figure 6B, C). The visualizations show that enrichment of abundant sequences occurred in round 3. Walton *et al.* showed that the most abundant ligase ribozymes possessed a conserved sequence motif—‘5’-UGCGG-3′ that likely helps these ribozymes bind their RNA substrate—5′ ACCACCGCAUUCCGCA-3′ via complementary base-pairing between the underlined nucleotides (18). We used the *Sequence Motif Tracker* tool to generate spreadsheets containing sequences that possess this motif (‘TGCGG’) and a plot illustrating how these sequences were enriched across selection. Significant enrichment was observed in round 3, which coincided with the enrichment of the most abundant sequences (Figure 6C, D), underscoring the importance of this motif in ribozyme function. The test dataset was not suited for the *Mutational Intermediates Finder* tool and therefore this tool was not tested here.

After completing the first pipeline, we returned to the last page of the pre-processing module (*Step 5*) and selected the *Analyze Sequence Clusters* pipeline in *Step 5d* after re-entering input parameters in *Steps 5b and 5c*. This pipeline allows immediate access to spreadsheets with sequence clusters from each round and the identities of their peak sequences (Figure 7A). To test the *Conservation Finder* tool, we analyzed the most abundant cluster. Certain positions like 6, 16, 10, 20, 25 and 37 were found to be more



Selection Statistics

Upload Times: \Choose Rounds(0.419)\Uploa

Round 1: Selection Statistics	
Number of Total Sequences	100000
Number of High Quality Sequences	48942
Number of Unique Sequences	48481
Fraction Unique Sequences	0.991

Round 2: Selection Statistics	
Number of Total Sequences	100000
Number of High Quality Sequences	51123
Number of Unique Sequences	50724
Fraction Unique Sequences	0.992

Round 3: Selection Statistics	
Number of Total Sequences	100000
Number of High Quality Sequences	51215
Number of Unique Sequences	9772
Fraction Unique Sequences	0.191

Round 4: Selection Statistics	
Number of Total Sequences	100000
Number of High Quality Sequences	49385
Number of Unique Sequences	7319
Fraction Unique Sequences	0.148

Top Sequence Tracker

Number of Sequences to Track = 5

Quality Cutoff = 98

Top Sequence Finder

Upload Times: \Choose Rounds(0.57)\Upload(14.332)\Upload(15.418)\Upload(10.899)\Upload(26.584)\Quality Filter(53.541)\Preview(61.48)\Selection statistics(3.477)\Most Abundant(61.559)

Number of Most Abundant Sequences to Identify = 5

Quality Cutoff = 98

Total Number of High Quality Sequences = 48942

Most Abundant Sequences [sequence, reads]=

```
[['GAATGCTGCCAACCGTGCGGGCTAATTGGCAGACTGAGCT', 12344], ['CCTAGCTAGCGCTGACTAGGACAGATGAGCGGCGGAACCA', 3255],
['TTAGTGAAATTGGTGCCCAAGCAGAGAATTGGGATAAATC', 2230], ['GAACCCTTATCACAGTCGTGCGGATTTGTAAGCCTAAGCG', 1390],
['TATTGGGTGCCGATTTCCTCTAATTTTAAAGTGCGGATAA', 1165]]
```

Sequence Motif Tracker

Upload Times: \Choose Rounds(0.57)\Upload(14.332)\Upload(15.418)\Upload(10.899)\Upload(26.584)\Q
Abundant(61.559)\Track Seqs(65.99)\Motif(2.352)\Motif(3.138)\Motif(2.443)

Region(s) to Analyze Emergence = ['TGCGG']

Quality Cutoff = 98

Figure 6. Output from the *Analyze Individual Sequences* pipeline using sequence data from *in vitro* selection of ligase ribozymes (18). (**A**) Selection Statistics summarizes selection outcomes for each round by calculating the total high-quality sequences and the fraction of unique sequences, which indicates sequence enrichment. (**B**) Top Sequence Tracker plots the relative abundances of the five most abundant sequences across selection. This tool reveals that the most abundant sequence covered ∼25% of all sequence population and sequence enrichment was pronounced in round 3. (**C**) Top Sequence Finder identifies and lists the five most abundant sequences with their read counts. (**D**) Sequence Motif Tracker lists all sequences containing the user-defined sequence motif—'TGCGG' and graphically depicts their enrichment. The plot indicates that sequences containing the TGCGG sequence motif were significantly enriched in round 3, concurrent with overall enrichment.

conserved, while other positions like 4, 23, 24, 33, 38 were less conserved (Figure 7B). Most of these nucleotides have been shown to be important for function (unpublished results). Additionally, the heatmap revealed conservation for the 'TGCGG' motif for nucleotides 16–20, further validating its utility. Finally, the *Cluster Peak Tracker* tool was used to quantify the relative abundances of the five most abundant peak sequences along the selection trajectory. The most abundant peak sequences matched the reported peak sequences in Walton *et al.*, where clustering was performed by Clustal Omega (18).

Case study: computation speed evaluation

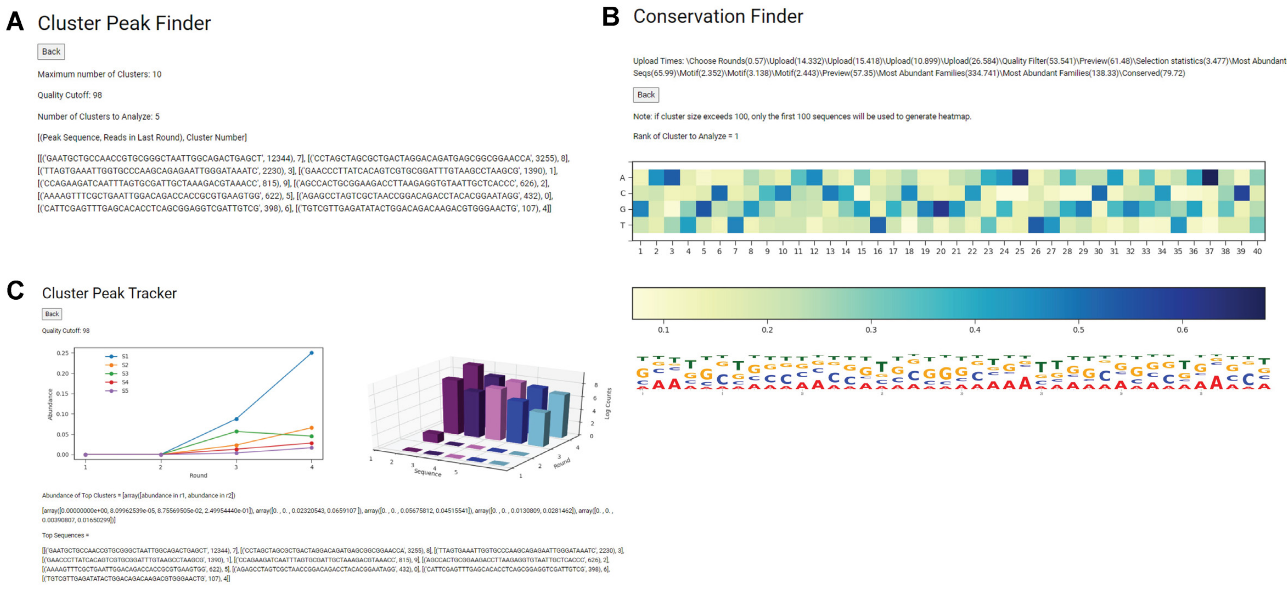
We measured the computation time for each tool with either 10 000 or 100 000 sequences per sequence file (on Chrome browser). For 10 000 sequences, file upload and pre-processing steps were completed in 2 min, whereas it took 4 min for 100 000 sequences. The *Analyze Individual Sequences* and *Analyze Sequence Clusters* pipeline took 40 and 170 s for 10 000 sequences, respectively. For 100 000 sequences, these pipelines took 6.5 and 12 min, respectively. Additionally, we timed each step when REVERSE was used to analyze all sequences from a single file that had been quality filtered by the Galaxy server. The pre-processing module (without quality filter) took 1.2 min, and the *Analyze Individual Sequences* and *Analyze Sequence Clusters* pipeline took 1.5 and 7.5 min, respectively. A detailed breakdown for each tool in each instance is provided in Supplementary Table S2.

Case study: applicability of analyzing a subset of sequence data

As analyzing a subset of the total number of sequences significantly reduced computation times, we wondered if results obtained from the analysis of 10 000 sequences could be extrapolated to the analysis of more sequences. We compared results obtained by analyzing 10 000, 100 000 and all sequences using the sequence file from the final round of selection (Supplementary Table S3). The values for fraction of unique sequences, which indicates enrichment, were somewhat comparable between 10 000 (0.246) and 100 000 (0.148) sequences. The most abundant sequences were identical in all three cases and their relative abundances were nearly identical. These results suggest that analyzing a subset of the sequence data can provide similar outcomes as analyzing all sequences, while significantly accelerating the process.

DISCUSSION

We developed the REVERSE web server as a one-stop solution for the bioinformatics needs of analyzing next-generation sequencing data obtained from combinatorial selections. REVERSE is completely GUI-based, which allows experimentalists without programming skills or familiarity with the command-line to integrate NGS with their combinatorial selection experiments. The REVERSE workflow is intuitive and streamlined, where the user uploads multiple raw FASTQ files once, without the need to



Cluster Peak Finder

Maximum number of Clusters: 10

Quality Cutoff: 98

Number of Clusters to Analyze: 5

[(Peak Sequence, Reads in Last Round), Cluster Number]

```
[[('GAATGCTGCCAACCGTGCGGGCTAATTGGCAGACTGAGCT', 12344), 7], [('CCTAGCTAGCGCTGACTAGGACAGATGAGCGGCGGAACCA', 3255), 8],
[('TTAGTGAAATTGGTGCCCAAGCAGAGAATTGGGATAAATC', 2230), 3], [('GAACCCTTATCACAGTCGTGCGGATTTGTAAGCCTAAGCG', 1390), 1],
[('CCAGAAGATCAATTTAGTGCGATTGCTAAAGACGTAAACC', 815), 9], [('AGCCACTGCGGAAGACCTTAAGAGGTGTAATTGCTCACCC', 626), 2],
[('AAAAGTTTCGCTGAATTGGACAGACCACCGCGTGAAGTGG', 622), 5], [('AGAGCCTAGTCGCTAACCGGACAGACCTACACGGAATAGG', 432), 0],
[('CATTCGAGTTTGAGCACACCTCAGCGGAGGTCGATTGTCG', 398), 6], [('TGTCGTTGAGATATACTGGACAGACAAGACGTGGGAACTG', 107), 4]]
```

Conservation Finder

Upload Times: \Choose Rounds(0.57)\Upload(14.332)\Upload(15.418)\Upload(10.899)\Upload(26.584)\Quality Filter(53.541)\Preview(61.48)\Selection statistics(3.477)\Most Abundant(61.559)\Track Seqs(65.99)\Motif(2.352)\Motif(3.138)\Motif(2.443)\Preview(57.35)\Most Abundant Families(334.741)\Most Abundant Families(138.33)\Conserved(79.72)

Note: if cluster size exceeds 100, only the first 100 sequences will be used to generate heatmap.

Rank of Cluster to Analyze = 1

Cluster Peak Tracker

Quality Cutoff: 98

Abundance of Top Clusters = [array([abundance in r1, abundance in r2])

```
[array([0.00000000e+00, 8.09962539e-05, 8.75569505e-02, 2.49954440e-01]), array([0. , 0. , 0.02320543, 0.0659107 ]), array([0. , 0. , 0.05675812, 0.04515541]), array([0. , 0. , 0.0130809, 0.0281462]), array([0. , 0. ,
0.00390807, 0.01650299])]
```

Top Sequences =

```
[[('GAATGCTGCCAACCGTGCGGGCTAATTGGCAGACTGAGCT', 12344), 7], [('CCTAGCTAGCGCTGACTAGGACAGATGAGCGGCGGAACCA', 3255), 8], [('TTAGTGAAATTGGTGCCCAAGCAGAGAATTGGGATAAATC', 2230), 3],
[('GAACCCTTATCACAGTCGTGCGGATTTGTAAGCCTAAGCG', 1390), 1], [('CCAGAAGATCAATTTAGTGCGATTGCTAAAGACGTAAACC', 815), 9], [('AGCCACTGCGGAAGACCTTAAGAGGTGTAATTGCTCACCC', 626), 2],
[('AAAAGTTTCGCTGAATTGGACAGACCACCGCGTGAAGTGG', 622), 5], [('AGAGCCTAGTCGCTAACCGGACAGACCTACACGGAATAGG', 432), 0], [('CATTCGAGTTTGAGCACACCTCAGCGGAGGTCGATTGTCG', 398), 6],
[('TGTCGTTGAGATATACTGGACAGACAAGACGTGGGAACTG', 107), 4]]
```

Figure 7. Output from the *Analyze Sequence Clusters* pipeline using sequence data from *in vitro* selection of ligase ribozymes (18). (**A**) Cluster Peak Finder identifies the peak sequences in each of the five most abundant clusters with their cluster assignment and read counts. Identification of peak sequences from distinct clusters allows wider sampling of selected sequence space than identifying the most abundant sequences. (**B**) Conservation Finder quantifies nucleotide conservation at each position in sequences of a user-defined cluster and depicts the results in a customizable heatmap. It also generates a sequence logo depicting a consensus sequence for the cluster from conservation data. (**C**) Cluster Peak Tracker plots the enrichment of peak sequences across selection and captures the population dynamics of the five most abundant clusters across selection.

use processed files for subsequent steps. REVERSE identifies the most abundant sequences/sequence clusters, graphically tracks the relative abundances of selected sequences across rounds of selection, and quantifies sequence enrichment in minutes. By removing the need for writing custom codes and/or having to rely on bioinformatic collaborations, we think REVERSE will reduce the time lag between NGS data acquisition and biochemical characterization of isolated sequences. In addition to generating the results common to most combinatorial selections, REVERSE performs more advanced computations such as conservation analysis and identification of mutational intermediates between source and target sequences. Although REVERSE was developed for analyzing nucleic acid selection/evolution experiments, it can be used to analyze data generated by other experiments that use combinatorial libraries.

Web-based bioinformatic platforms for analyzing large NGS data files require non-trivial computational resources and are time-intensive. By allowing users to analyze a representative subset of their data, REVERSE can significantly reduce analysis times. For example, the pre-processing module is 2-fold faster for the same dataset if 10 000 sequences were analyzed instead of 100 000 (Supplementary Table S3) without greatly compromising accuracy. Although a subset of the data is likely sufficient for the most basic combinatorial selections needs, certain tasks require greater coverage of sequence data. These include detection of enrichment in earlier rounds, conservation analysis of specific sequence clusters, discovering neutral networks between two distinct RNAs, and broader investigations into the architecture of evolutionary fitness landscapes. Due to the of '3GB per process' RAM limit imposed by PythonAnywhere, the host server, REVERSE cannot analyze all sequences from multiple sequence files simultaneously. With increase in traffic, we plan to host REVERSE in a private server within or outside PythonAnywhere which will remove the current memory limits. Currently, if users wish to analyze all sequences (and not a subset), they may upload pre-processed files from the Galaxy server to reduce the number of sequences in each file and consequently reduce computational demands. Alternatively, users may analyze sequence data from the final round if their primary goal is to identify the most abundant sequences/cluster peak sequences.

We are also working on implementing computational strategies to fragment sequence files into readily analyzable blocks (∼100 000 sequences), which will undergo parallel pre-processing before being concatenated for the analysis module (36). This strategy will be utilized for tools in the analysis module that benefit from greater sequence coverage. Once REVERSE is able to handle multiple full-length files, we will incorporate a paired-end read merger tool, which will reduce sequencing errors and thus increase high-quality sequence reads. In future iterations, we also plan to integrate the Infernal (37) secondary structure motif search to complement REVERSE's existing sequence motif search tool and incorporate the R-scape algorithm for covariation analysis (38). Since NGS can identify multiple variants of peak sequences within a cluster, multiple sequence alignments can be used to search for recurring structural motifs using Infernal and generate secondary structure models by covariation analysis using the R-scape algorithm. We hope that REVERSE will allow nonbioinformatician-experimentalists to benefit from NGS technologies in their *in vitro* selection/evolution experiments.

DATA AVAILABILITY

REVERSE is freely available at https://www.reverseserver.org/ and does not have a login requirement. The code for REVERSE is open source and available in the GitHub repository (https://github.com/szostaklab/REVERSE).

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank Dr Daniel Duzdevich for helpful comments on this project and thank Dr Jack W. Szostak for valuable suggestions throughout this project, critical reading of the manuscript, and funding. We also thank the reviewers for their valuable insights that helped us improve the web server and the accompanying manuscript.

FUNDING

Harvard Origins of Life Initiative Summer Fellowship; Harvard PRISE Fellowship; SETI Forward (to Z.W.). Funding for open access charge: MGH internal funding (to Dr Jack W. Szostak).
Conflict of interest statement. None declared.

REFERENCES

1. Tuerk,C. and Gold,L. (1990) Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. *Science (New York, N.Y.)*, **249**, 505–510.
2. Robertson,D.L. and Joyce,G.F. (1990) Selection in vitro of an RNA enzyme that specifically cleaves single-stranded DNA. *Nature*, **344**, 467–468.
3. Ellington,A.D. and Szostak,J.W. (1990) In vitro selection of RNA molecules that bind specific ligands. *Nature*, **346**, 818–822.
4. Li,L., Xu,S., Yan,H., Li,X., Yazd,H.S., Li,X., Huang,T., Cui,C., Jiang,J. and Tan,W. (2021) Nucleic acid aptamers for molecular diagnostics and therapeutics: advances and perspectives. *Angew. Chem. Int. Ed. Engl.*, **60**, 2221–2231.
5. Ma,L. and Liu,J. (2020) Catalytic nucleic acids: biochemistry, chemical biology, biosensors, and nanotechnology. *Iscience*, **23**, 100815.
6. Micura,R. and Höbartner,C. (2020) Fundamental studies of functional nucleic acids: aptamers, riboswitches, ribozymes and DNAzymes. *Chem. Soc. Rev.*, **49**, 7331–7353.
7. Bendixsen,D.P., Collet,J., Østman,B. and Hayden,E.J. (2019) Genotype network intersections promote evolutionary innovation. *PLoS Biol.*, **17**, e3000300.
8. Peri,G., Gibard,C., Shults,N.H., Crossin,K. and Hayden,E.J. (2022) Dynamic RNA fitness landscapes of a group I ribozyme during changes to the experimental environment. *Mol. Biol. Evol.*, **39**, msab373 .
9. Athavale,S.S., Spicer,B. and Chen,I.A. (2014) Experimental fitness landscapes to understand the molecular evolution of RNA-based life. *Curr. Opin. Chem. Biol.*, **22**, 35–39.
10. Blanco,C., Janzen,E., Pressman,A., Saha,R. and Chen,I.A. (2019) Molecular fitness landscapes from high-coverage sequence profiling. *Annu. Rev. Biophys.*, **48**, 1–18.
11. Jimenez,J.I., Xulvi-Brunet,R., Campbell,G.W., Turk-MacLeod,R. and Chen,I.A. (2013) Comprehensive experimental fitness landscape and evolutionary network for small RNA. *Proc. Nat. Acad. Sci. U.S.A.*, **110**, 14984–14989.
12. Pressman,A., Moretti,J.E., Campbell,G.W., Muller,U.F. and Chen,I.A. (2017) Analysis of in vitro evolution reveals the underlying distribution of catalytic activity among random sequences. *Nucleic Acids Res.*, **45**, 10922.
13. Pressman,A.D., Liu,Z., Janzen,E., Blanco,C., Muller,U.F., Joyce,G.F., Pascal,R. and Chen,I.A. (2019) Mapping a systematic ribozyme fitness landscape reveals a frustrated evolutionary network for self-aminoacylating RNA. *J. Am. Chem. Soc.*, **141**, 6213–6223.
14. Xulvi-Brunet,R., Campbell,G.W., Rajamani,S., Jimenez,J.I. and Chen,I.A. (2016) Computational analysis of fitness landscapes and evolutionary networks from in vitro evolution experiments. *Methods*, **106**, 86–96.
15. Chen,X., Li,N. and Ellington,A.D. (2007) Ribozyme catalysis of metabolism in the RNA world. *Chem. Biodivers.*, **4**, 633–655.
16. Hirao,I. and Ellington,A.D. (1995) Re-creating the RNA world. *Curr. Biol.*, **5**, 1017–1022.
17. Müller,U.F. (2006) Re-creating an RNA world. *Cell. Mol. Life Sci.*, **63**, 1278–1293.
18. Walton,T., DasGupta,S., Duzdevich,D., Oh,S.S. and Szostak,J.W. (2020) In vitro selection of ribozyme ligases that use prebiotically plausible 2-aminoimidazole-activated substrates. *Proc. Nat. Acad. Sci. U.S.A.*, **117**, 5741–5748.
19. Newton,M.S., Cabezas-Perusse,Y., Tong,C.L. and Seelig,B. (2020) In vitro selection of peptides and proteins-advantages of mRNA display. *ACS Synth. Biol.*, **9**, 181–190.
20. Christiansen,A., Kringelum,J.V., Hansen,C.S., Bogh,K.L., Sullivan,E., Patel,J., Rigby,N.M., Eiwegger,T., Szepfalusi,Z., de Masi,F. *et al.* (2015) High-throughput sequencing enhanced phage display enables the identification of patient-specific epitope motifs in serum. *Sci. Rep.*, **5**, 12913.
21. Yokobayashi,Y. (2019) Applications of high-throughput sequencing to analyze and engineer ribozymes. *Methods*, **161**, 41–45.
22. Huynen,M.A., Stadler,P.F. and Fontana,W. (1996) Smoothness within ruggedness: the role of neutrality in adaptation. *Proc. Nat. Acad. Sci. U.S.A.*, **93**, 397–401.
23. Schultes,E.A. and Bartel,D.P. (2000) One sequence, two ribozymes: implications for the emergence of new ribozyme folds. *Science (New York, N.Y.)*, **289**, 448–452.
24. DasGupta,S., Nykiel,K. and Piccirilli,J.A. (2021) The hammerhead self-cleaving motif as a precursor to complex endonucleolytic ribozymes. *RNA*, **27**, 1017–1024.
25. Ge,S.X., Son,E.W. and Yao,R. (2018) iDEP: an integrated web application for differential expression and pathway analysis of RNA-Seq data. *BMC Bioinf.*, **19**, 534.
26. Cheng,X., Yan,J., Liu,Y., Wang,J. and Taubert,S. (2021) eVITTA: a web-based visualization and inference toolbox for transcriptome analysis. *Nucleic Acids Res.*, **49**, W207–W215.
27. Kechin,A., Boyarskikh,U., Kel,A. and Filipenko,M. (2017) cutPrimers: a new tool for accurate cutting of primers from reads of targeted next generation sequencing. *J. Comput. Biol.*, **24**, 1138–1143.
28. Martin,M. (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. J.***17**, 3.
29. Bolger,A.M., Lohse,M. and Usadel,B. (2014) Trimmomatic: a flexible trimmer for illumina sequence data. *Bioinformatics*, **30**, 2114–2120.
30. Masella,A.P., Bartram,A.K., Truszkowski,J.M., Brown,D.G. and Neufeld,J.D. (2012) PANDAseq: paired-end assembler for illumina sequences. *BMC Bioinf.*, **13**, 31.
31. Zhang,J., Kobert,K., Flouri,T. and Stamatakis,A. (2014) PEAR: a fast and accurate illumina paired-end reAd mergeR. *Bioinformatics*, **30**, 614–620.
32. Blanco,C., Verbanic,S., Seelig,B. and Chen,I.A. (2020) EasyDIVER: a pipeline for assembling and counting high-throughput sequencing data from in vitro evolution of nucleic acids or peptides. *J. Mol. Evol.*, **88**, 477–481.
33. Alam,K.K., Chang,J.L. and Burke,D.H. (2015) FASTAptamer: a bioinformatic toolkit for High-throughput sequence analysis of combinatorial selections. *Mol. Ther. Nucleic Acids*, **4**, e230.
34. Hoinka,J., Backofen,R. and Przytycka,T.M. (2018) AptaSUITE: a full-featured bioinformatics framework for the comprehensive

analysis of aptamers from HT-SELEX experiments. *Mol. Ther. Nucleic Acids*, **11**, 515–517.

35. Hoinka,J., Dao,P. and Przytycka,T.M. (2015) AptaGUI-A graphical user interface for the efficient analysis of HT-SELEX data. *Mol. Ther. Nucleic Acids*, **4**, e257.
36. Duzdevich,D., Carr,C.E. and Szostak,J.W. (2020) Deep sequencing of non-enzymatic RNA primer extension. *Nucleic Acids Res.*, **48**, e70.
37. Nawrocki,E.P. and Eddy,S.R. (2013) Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics*, **29**, 2933–2935.
38. Rivas,E., Clements,J. and Eddy,S.R. (2017) A statistical test for conserved RNA structure shows lack of evidence for structure in lncRNAs. *Nat. Methods*, **14**, 45–48.