

Supplementary Information

EasyDIVER+: an advanced tool for analyzing high throughput sequencing data from *in vitro* evolution of nucleic acids or amino acids

Celia Blanco^{1,2,†}, Allison Tee^{1,3}, Pramesh Sharma⁴, Matilda S. Newton^{5,6,7}, Kun-Hwa Lee^{5,6}, Samuel E. Erickson^{5,6}, Burckhard Seelig^{5,6} and Irene A. Chen².

1. Blue Marble Space Institute of Science, 600 1st Avenue, 1st Floor, Seattle, WA, USA

2. Department of Chemical and Biomolecular Engineering, University of California, Los Angeles, CA, USA

3. Department of Computer Science, Stanford University, CA, USA

4. College of Engineering, University of California, Davis, CA, USA

5. Department of Biochemistry, Molecular Biology and Biophysics, University of Minnesota, Minneapolis, MN, USA

6. BioTechnology Institute, University of Minnesota, St. Paul, MN, USA

7. Royal Society of New Zealand, Wellington, New Zealand

[†] to whom correspondence should be addressed: celia.blanco@bmsis.org

Easy Diver+

REQUIRED

Input Directory Path: ?

OPTIONAL

Output Directory Path: ?

☐ Skip Processing (enrichment analysis only) ?

Forward Primer Sequence Extraction: ?

Reverse Primer Sequence Extraction: ?

Extra Flags for PANDASeq (use quotes, e.g. "-L 50"): ?

☒ Translate to Amino Acids ?

Choose Genetic Code for Translation: ?

☐ Retain Individual Lane Outputs ?

☐ Use Local PANDASeq (Verify you have working version first!) ?

☒ Run Enrichment Analysis for Consecutive Rounds ?

Output Decimal Precision: ?

REQUIRED: Sort files into rounds and types!

Supplementary Figure 1. EasyDIVER+ Graphical User Interface for sequencing data processing and analysis.

File Sorting Interface

How many rounds?

test_Neg_S1

test_Neg_S2

test_Out_S1

test_Out_S2

test_in_S1

test_in_S2

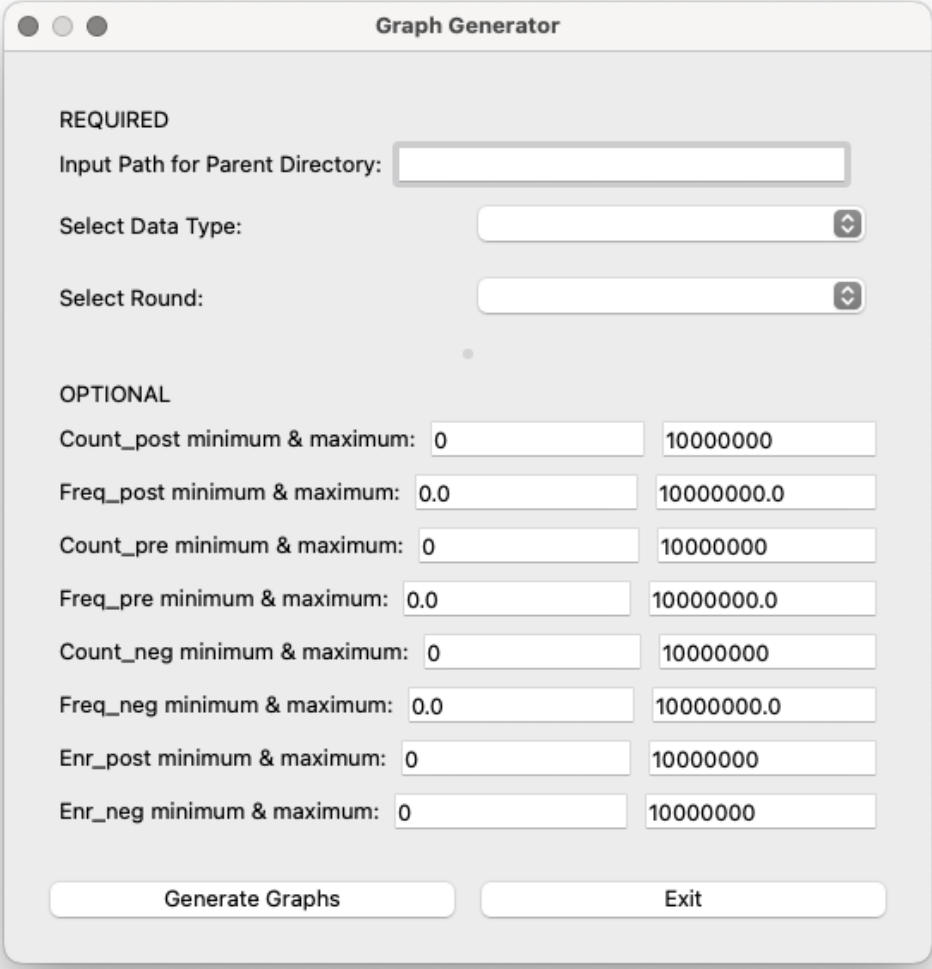
Round 1

Pre: Post: Neg:

Round 2

Pre: Post: Neg:

Supplementary Figure 2. EasyDIVER+ Graphical User Interface for assigning sample files into buckets (pre-selection, post-selection, and negative control)



The image shows a software window titled "Graph Generator". It contains two main sections: "REQUIRED" and "OPTIONAL".

REQUIRED

- Input Path for Parent Directory:** A text input field.
- Select Data Type:** A dropdown menu.
- Select Round:** A dropdown menu.

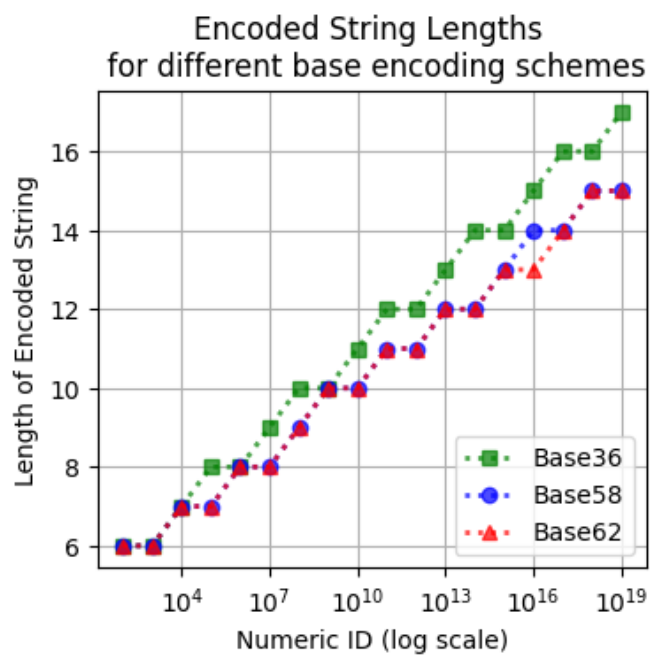
OPTIONAL

Below the optional section, there is a small dot separator.

Count_post minimum & maximum:	0	10000000
Freq_post minimum & maximum:	0.0	10000000.0
Count_pre minimum & maximum:	0	10000000
Freq_pre minimum & maximum:	0.0	10000000.0
Count_neg minimum & maximum:	0	10000000
Freq_neg minimum & maximum:	0.0	10000000.0
Enr_post minimum & maximum:	0	10000000
Enr_neg minimum & maximum:	0	10000000

At the bottom of the window, there are two buttons: "Generate Graphs" and "Exit".

Supplementary Figure 3. EasyDIVER+ Graph Generator interface.



Supplementary Figure 4. Length of unique names across different encoding schemes: Base36, Base58, and Base62. Base58 strikes a balance by offering short, clear identifiers while avoiding ambiguous characters

File/Directory	Description
counts_nt	DNA counts files for every sample (unique sequences, read counts, and frequency)
counts_aa	Amino acid counts files for every sample (unique sequences, read counts, and frequency)
fastas	Joined FASTA files
fastqs	Joined FASTQ files
histos	DNA sequence length distributions and peptide sequence length distribution (if translation was chosen)
log_output_[date-time].txt	Log file with information on parameters used and outcome of each processing step
analysis_output_nt	DNA enrichment analysis output files (one per round), and frequency and enrichment output files for all sequences across all rounds
analysis_output_aa	Amino acid enrichment analysis output files (one per round), and frequency and enrichment output files for all sequences across all rounds
individual.lanes	All processing files (joined FASTA files, joined FASTQ files, text counts files, and text histograms) corresponding to individual lanes (if requested using the flag -r)
log_easydiverplus_[date-time].txt	Log file with full terminal output and output directory tree (files and sizes).

Supplementary Table 1. Overview of output files generated by EasyDIVER+ during the data processing stage. The table lists the different file types and directories created, providing essential information for subsequent analysis steps.

Column Name	Description
Unique_Sequence_Name	Base58-encoded unique sequence name.
Sequence	DNA or amino acid sequence.
Count_pre	Read count in pre-selection.
Count_Lower_pre	Lower 95% confidence interval for Count_pre.
Count_Upper_pre	Upper 95% confidence interval for Count_pre.
Freq_pre	Frequency (relative abundance) in pre-selection.
Freq_Lower_pre	Lower 95% confidence interval for Freq_pre.
Freq_Upper_pre	Upper 95% confidence interval for Freq_pre.
Count_post	Read count in post-selection.
Count_Lower_post	Lower 95% confidence interval for Count_post.
Count_Upper_post	Upper 95% confidence interval for Count_post.
Freq_post	Frequency (relative abundance) in post-selection.
Freq_Lower_post	Lower 95% confidence interval for Freq_post.
Freq_Upper_post	Upper 95% confidence interval for Freq_post.
Count_neg	Read count in negative control.
Count_Lower_neg	Lower 95% confidence interval for Count_neg.
Count_Upper_neg	Upper 95% confidence interval for Count_neg.
Freq_neg	Frequency (relative abundance) in negative control.
Freq_Lower_neg	Lower 95% confidence interval for Freq_neg.
Freq_Upper_neg	Upper 95% confidence interval for Freq_neg.
Enr_post	Enrichment in post-selection ($\text{Freq_post}/\text{Freq_pre}$).
Enr_post_lower	Lower 95% confidence interval for Enr_post.
Enr_post_upper	Upper 95% confidence interval for Enr_post.
Enr_neg	Enrichment in negative control ($\text{Freq_neg}/\text{Freq_pre}$).
Enr_neg_lower	Lower 95% confidence interval for Enr_neg.
Enr_neg_upper	Upper 95% confidence interval for Enr_neg.
Enr_ratio	Relative enrichment ($\text{Enr_post}/\text{Enr_neg}$).
Enr_ratio_lower	Lower 95% confidence interval for Enr_ratio.
Enr_ratio_upper	Upper 95% confidence interval for Enr_ratio.

Supplementary Table 2. Description of columns in the *round_00X_enrichment_analysis.csv* files generated by EasyDIVER+. These metrics include read counts, frequencies (relative

abundances), and enrichment values for each sequence across pre-selection, post-selection, and negative control samples, along with their associated 95% confidence intervals.

File name	Description
all_rounds_enrichment_negative_results.csv *	Enrichment in the negative control selection for each unique sequence across all rounds of selection.
all_rounds_enrichment_post_results.csv	Enrichment in the post-selection for each unique sequence across all rounds of selection.
all_rounds_enrichment_ratio_results.csv *	Relative enrichment for each unique sequence across all rounds of selection.
all_rounds_frequency_negative_results.csv *	Frequency (relative abundance) in the negative control selection for each unique sequence across all rounds of selection.
all_rounds_frequency_post_results.csv	Frequency (relative abundance) in the post-selection for each unique sequence across all rounds of selection.
all_rounds_frequency_pre_results.csv	Frequency (relative abundance) in the pre-selection for each unique sequence across all rounds of selection.

Supplementary Table 3. Summary of the additional output files generated by EasyDIVER+ analysis, with data for enrichment and frequency metrics for each unique sequence across all rounds of selection, as defined in the main text (section 'Enrichment analysis'). Files are stored in the *analysis_outputs* directory. Asterisk (*) indicates files that can only be generated if negative control data is provided for at least one round in the selection.