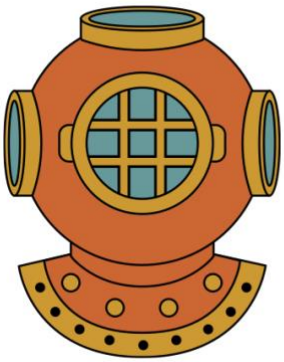


EasyDIVER

Easy pre-processing and **D**ereplication of **I**n **V**itro **E**volution **R**eads

User's guide
Version 2.0
(May 2020)

Celia Blanco[1,2], Samuel Verbanic[2,3], Burckhard Seelig[4,5] and Irene A. Chen[1,2,3]

1. Department of Chemistry and Biochemistry 9510, University of California, Santa Barbara, CA 93106
2. Department of Chemical and Biomolecular Engineering, University of California, Los Angeles, CA 90095
3. Program in Biomolecular Sciences and Engineering, University of California, Santa Barbara, CA 93106
4. Department of Biochemistry, Molecular Biology and Biophysics, University of Minnesota, Minneapolis, MN, 55455, USA
5. BioTechnology Institute, University of Minnesota, St. Paul, MN, 55108, USA

If you use Easydiver, please cite the paper:

Celia Blanco*, Samuel Verbanic*, Burckhard Seelig and Irene A. Chen. EasyDIVER: a pipeline for assembling and counting high throughput sequencing data from in vitro evolution of nucleic acids or peptides. *Under review*.

For feedback, suggestions, technical support, etc., please email us at celiablanco@ucla.edu or verbanic@ucla.edu.

When reporting bugs, please include the full output (error messages included) printed in the terminal when running the pipeline.

DESCRIPTION

EasyDIVER converts raw, paired-end, demultiplexed Illumina read files into processed, dereplicated data ready for analysis. The algorithm performs the following:

1. Joins the raw data using PANDAseq.
2. Extracts the insert sequence based on user-supplied primer sequences.
3. Optionally translates into amino acids.
4. Generates counts files.
5. Collects sequence length distributions (histos).
6. Creates a log file.

SYNOPSIS

```
easydiver.sh -i input.directory [-o output.directory] [-p
forward.primer] [-q reverse.primer] [-a] [-r] [-T threads] [-e] [-h]
```

Alternatively, a more user-friendly (but less versatile) solution can be optionally used. If no flags are provided the user will be prompted for input values in the command line in verbose form. To use the prompted input version, run:

```
easydiver.sh
```

OPTIONS

-i input.directory

Required. Input directory path and name. If no value is provided, an error message will be printed in the terminal: `ERROR: No input filepath supplied` and no further action will be performed.

-o output.directory

Optional. Output directory path and name. If no value is provided, the default value `/pipeline.output` will be used.

-p forward.primer

Optional. Extraction forward DNA primer. If a forward primer sequence is provided, the pipeline strips out the primer from the start of the sequence. Any sequence before the provided primer will be discarded.

-q reverse.primer

Optional. Extraction reverse DNA primer. If a reverse primer sequence is provided, the pipeline strips out the primer at the start of the sequence. Any sequence after the provided primer will be discarded.

-a

Optional. Translation into amino acids is performed. DNA sequences are translated using the standard genetic code, and the resulting sequences are dereplicated. By default, translation is not performed.

-r

Optional. Files for individual lanes are retained. By default, the script will suppress outputs from individual lanes.

-T threads

Optional. Number of threads used for computation. Default value is 1.
The number of threads that may be used is dependent on the user's CPU (for example, if using a machine with 16 threads, 14 could be a desirable number). The default value of 1 would be suboptimal for multi-core machines.

-e

Optional. Additional internal PANDAseq flags. Values must be entered in quotation marks (e.g. `-e "-L 50"`). Default value is `"-l 1 -d rbfkms"`. For more information see the PANDAseq manual.

-h

If used, a help message will be printed in the terminal and no further action will be performed.

USAGE

The pipeline (easydiver.sh) and the translator (translator.py) are available in the GitHub repository:

https://github.com/ichen-lab-ucsb/EasyDIVER

Both files must be placed in the same location. For simplicity, the instructions below assume both files are located in a directory that is in the user's PATH environment variable upon download. For example, for Unix/Linux users, scripts could be placed in `/usr/local/bin/`. If this is not the case, the path to the file location should be provided when calling it.

To make EasyDIVER executable, enter the following command from the local directory where it's stored:

```
chmod +x easydiver.sh
```

EXAMPLE

To test the pipeline and troubleshoot potential issues, we provide two samples of test data in the GitHub repository. An example of command to run the pipeline using the test data provided in the GitHub repository:

```
easydiver.sh -i ./ -o ./output -p GGCGGAAAGCACATCTGC -T 14 -a -r
```

Note: If the file EasyDIVER.sh is not made executable, the command bash must be used (e.g. `bash easydiver.sh -i [-o -p -q -h -a -r -T -e]`).

Alternatively, the prompted input version can be requested using the flag -v:

```
easydiver.sh -v
```

INPUT

All input files must be:

1. Located in the same directory (even reads from separate lanes).
2. In FASTQ format
3. Named using the standard Illumina naming scheme: `sample-name_S#_L00#_R#_001.fastq`
4. In either .fastq or .fastq.gz extensions.

OUTPUT

For each sample, the pipeline combines the reads from every lane, and redirects the outputs to the following sub-directories:

fastqs contains the joined fastq files
fastas contains the joined fasta files
counts contains DNA counts files for every sample
counts.aa contains peptide count files for every sample (if translation is requested using the flag `-a`)
histos contains the DNA sequence length distributions and the peptide sequence length distribution (if translation is required)
individual.lanes contains the files (joined fasta files joined fastq files, text counts files and text histograms) corresponding to the individual lanes (if requested using the flag `-r`)

If translation to amino acids is desired (indicated by the use of the flag `-a`) the counts files are translated using the standard genetic code, and the resulting sequences are dereplicated. Count files and length distributions are created for the amino acid sequences as well. All sequence length distributions are redirected to the directory **histos**.

By default, the script will suppress outputs from individual lanes. If you wish to retain the individual lane outputs, use the `-r` flag. If the flag `-r` is used, files corresponding to the individual lanes (joined fasta files joined fastq files, text counts files and text histograms) are retained and redirected to the subdirectory called **individual.lanes**.

For the record, and to monitor the success of the run, a single log text file with the parameters used and the number of sequences in the fastq and counts files is created at the end of the process.

After running the pipeline for the example above, the terminal output will look like this:

```
raw.reads username$ easydiver.sh -i ./ -o ./output -p GGCGGAAAGCACATCTGC -T 14 -a -r
 _______                 _____ ________      ________ _____
|  ____|               |  __ \_    _\ \    / /  ____|  __ \
| |__   ___  ___ _   _| |  | || |  \ \  / /| |__  | |__) |
|  __| / _ \/ __| | | | |  | || |   \ \/ / |  __| |  _  /
| |___| (_| \__ \ |_| | |__| || |_   \  /  | |____| | \ \
|______\__,_|___/\__, |_____/_____|   \/   |______|_|  \_\
                  __/ |
                 |___/
+-------------------------------------------------------------------------------+
| Thu Jan 30 13:06:22 PST 2020                                                  |
|                                                                               |
| Welcome to the pipeline for Easy pre-processing and Dereplication of In Vitro |
| Evolution Reads                                                               |
+-------------------------------------------------------------------------------+


-----Input directory path: /Users/username/Desktop/raw.reads

-----Output directory path: /Users/username/Desktop/raw.reads/output

-----Forward Primer: GGCGGAAAGCACATCTGC

-----No reverse primer supplied. Extraction will be skipped.

-----Number of threads = 14

-----No additional PANDAseq flags supplied.

-----Translation needed.

-----Individual lane outputs will be retained.


Input filecheck passed

Joining test1_S1_L001 reads & extracting primer...
Converting joined test1_S1_L001 FASTQ to FASTA...
Adding test1_S1_L001 reads to total test1_S1 reads...
Generating test1_S1_L001 nt length distribution for individual lanes...
Calculating unique & total reads for lane test1_S1_L001...
Collecting unique, total and sequences in file...

Joining test2_S2_L001 reads & extracting primer...
Converting joined test2_S2_L001 FASTQ to FASTA...
Adding test2_S2_L001 reads to total test2_S2 reads...
Generating test2_S2_L001 nt length distribution for individual lanes...
Calculating unique & total reads for lane test2_S2_L001...
Collecting unique, total and sequences in file...

Calculating unique & total reads for test1_S1...
Calculating unique & total reads for test2_S2...

Individual lane outputs will be retained

Generating test1_S1 DNA length distribution...
Translating test1_S1 DNA to peptides...
Generating test1_S1 aa length distribution...

Generating test2_S2 DNA length distribution...
Translating test2_S2 DNA to peptides...
Generating test2_S2 aa length distribution...

Run time: 87
```

If the prompted input version is requested (flag -v), the terminal output will look like this instead:

```
raw.reads username$ easydiver.sh -v
 _______                 _____ ________      _________ _____
|  ____|               |  __ \_   _\ \    / /  ____|  __ \
| |__   ___  ___ _   _| |  | || |  \ \  / /| |__  | |__) |
|  __| / _ \/ __| | | | |  | || |   \ \/ / |  __| |  _  /
| |___| (_| \__ \ |_| | |__| || |_   \  /  | |____| | \ \
|______\__,_|___/\__, |_____/_____|   \/   |______|_|  \_\
                  __/ |
                 |___/
+-------------------------------------------------------------------------------+
| Thu May 22 17:11:15 PST 2020                                                  |
|                                                                               |
| Welcome to the pipeline for Easy pre-processing and Dereplication of In Vitro |
| Evolution Reads                                                               |
+-------------------------------------------------------------------------------+

PROMPTED INPUT VERSION REQUESTED

Path to your input directory:
./

Path to your output directory (default value /pipeline.output):
./output

Forward primer sequence for extraction:
GGCGGAAAGCACATCTGC

Reverse primer sequence for extraction:


Number of threads desired for computation (default value 1):
14

Extra flags (default value "-l 1 -d rbfkms" ; see manual):


Perform translation into amino acids? (yes / no)
yes

Retain output files for individual lanes? (yes / no)
Yes

-----Input directory path: /Users/username/Desktop/raw.reads
-----Output directory path: /Users/username/Desktop/raw.reads/output
-----Forward Primer: GGCGGAAAGCACATCTGC
-----No reverse primer supplied. Extraction will be skipped.
-----Number of threads = 14
-----No additional PANDAseq flags supplied.
-----Translation needed.
-----Individual lane outputs will be retained.
```

```
Input filecheck passed

Joining test1_S1_L001 reads & extracting primer...
Converting joined test1_S1_L001 FASTQ to FASTA...
Adding test1_S1_L001 reads to total test1_S1 reads...
Generating test1_S1_L001 nt length distribution for individual lanes...
Calculating unique & total reads for lane test1_S1_L001...
Collecting unique, total and sequences in file...

Joining test2_S2_L001 reads & extracting primer...
Converting joined test2_S2_L001 FASTQ to FASTA...
Adding test2_S2_L001 reads to total test2_S2 reads...
Generating test2_S2_L001 nt length distribution for individual lanes...
Calculating unique & total reads for lane test2_S2_L001...
Collecting unique, total and sequences in file...

Calculating unique & total reads for test1_S1...
Calculating unique & total reads for test2_S2...

Individual lane outputs will be retained

Generating test1_S1 DNA length distribution...
Translating test1_S1 DNA to peptides...
Generating test1_S1 aa length distribution...

Generating test2_S2 DNA length distribution...
Translating test2_S2 DNA to peptides...
Generating test2_S2 aa length distribution...

Run time: 79
```

The outcoming file log.txt, for the same example parameters:

-----Input directory path: /Users/username/Desktop/raw.reads

-----Output directory path: /Users/username/Desktop/raw.reads/output

-----Forward Primer: GGCGGAAAGCACATCTGC

-----Individual lane outputs suppressed

-----# of threads = 14

-----Translation on

-----No additional PANDAseq flags

sample	fastq_R1	fastq_R2	unique_nt	total_nt	recovered_nt(%)	unique_aa	total_aa	recovered_aa(%)
test1_S1	64516	64516	54168	55576	86.14%	38695	55556	86.11%
test2_S2	53541	53541	45131	46605	87.05%	31147	46593	87.02%

The first lines of the outcoming peptide count file for sample sub1_S1 (sub1_S1_counts.aa.txt) will look as follow:

```
number of unique sequences = 38695

total number of molecules = 55556


AICGDVVATADTKIQYDSCEGCKGFSKRTVRKDLTYTCRDYKDCECYHKCLDLCQYCRYQKALAMGMKREAVQEEVGSHHQHHHGGSMGMSGSGTGY      2189  3.940%

AICGDYISAVDTQSKNDSCEGCKGFFKRTVRKDLTYTCRDNKNCECYHFCLQNCQYCRYQKALAMGMKREAVQEEVGSHHHHHHGGSMGMSGSGTGY      847   1.525%

AICGDYISAVDTQSKNDSCEGCKGFFKRTVRKDLTYTCRDNKDCECYHFCLQNCQYCRYQKALAMGMKREAVQEEVGSHHQHHHGGSMGMSGSGTGY      682   1.228%

AICGDVVATADTKIQYDSCEGCKGFSKRTVRKDLTYTCRDYKDCESYHKCSDLCQYCRYQKALAMGMKREAVQEEVGSHHQHHHGGSMGMSGSGTGY      589   1.060%

AICGDVVATADTKIQYDSCEGCKGFSKRTVRKDLTYTCRDYKDCECYHKCLDLCQYCRYQKALAMGMKRKAVQEEVGSHHQHHHGGSMGMSGSGTGY      511   0.920%

AICGDYISAVDTQSKNDSCEGCKGFFKRTVRKDLTYTCRDNKNCECYHFCLQNCQYCRYQKALAMGMKREAVQEEVGSHHQHHHGGSMGMSGSGTGY      328   0.590%

AICGDVVATADTKIQYDSCEGCKGFSKRTVRKDLTYTCRDYKNCESYHKCLDLCQYCRYQKALAMGMKREAVQEEVGSHHQHHHGGSMGMSGSGTGY      267   0.481%

AICGDVVATADTKIQYDSCEGCKGFSKRTVRKDLTYTCRDYKDCECYHKCLDLCQYCRYQKALAMGMKREAVQEEVGSHHQPHHGGSMGMSGSGTGY      264   0.475%

AICGDYISAVDTQSKNDSCEGCKGFFKRTVRKDLTYTCRDNKDCECYHFCLQNCQYCRYQKALAMGMKREAVQEEVGSHHHHHHGGSMGMSGSGTGY      224   0.403%

AICGDYISAVDTQSKNDSCEGCKGFFKRTVRKDLTNTCRDNKNCECYHFCLQNCQYCRYQKALAMGMKREAVQEEVGSHHHHHHGGSMGMSGSGTGY      215   0.387%
```

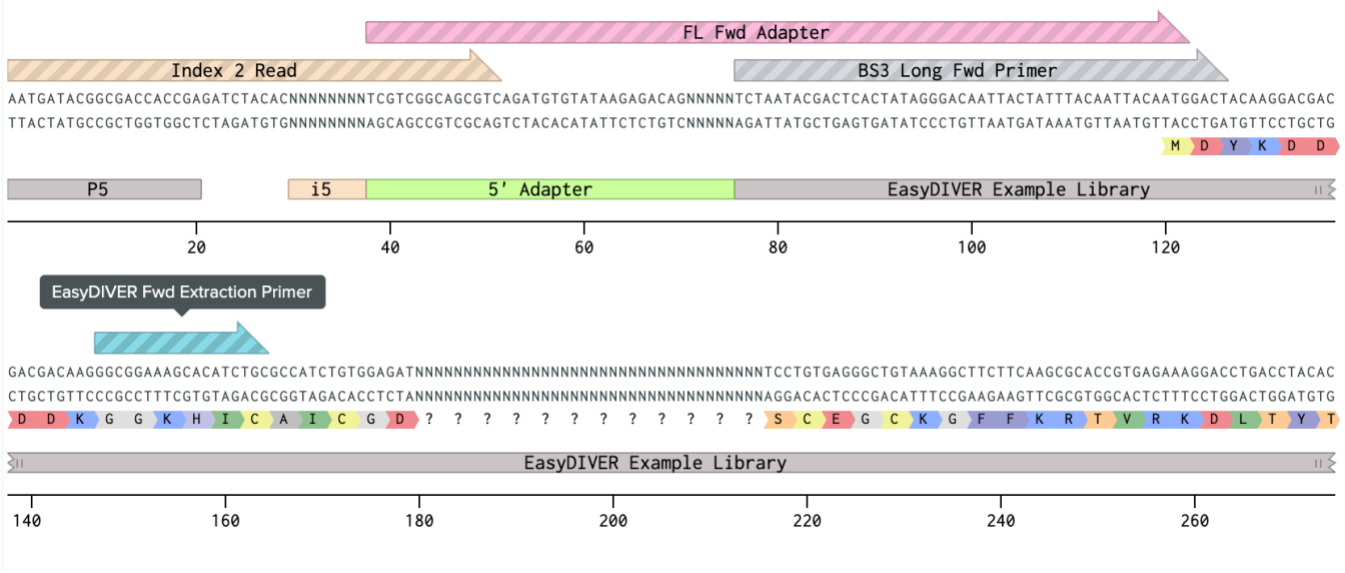
For more detailed information on EasyDIVER, please see our publication, or the README file available in the GitHub repository.

TIPS

To successfully process your data with EasyDIVER, the correct parameters must be selected. Here we will provide guidance on how to select parameters for your data.

- **Primer selection:** Your choice of primers will determine how your target sequences are extracted and set the reading frame for translation. You should select primers that will target conserved sequences in your library and place your extracted sequences in the desired reading frame.

    - **Example:** In the example command provided in this manual, the forward extraction primer sequence is entered as `-p GGCGGAAAGCACATCTGC`. Using the schematic below, we can see why this sequence was selected. It is in a semi-conserved portion of the library, and should be present in every sequence. The extracted sequence will be in our desired reading frame, starting with the amino acids `AICGD`, followed by the random portion of the library.



- **Threads:** Modern processors possess the capability to run processes in parallel by using multiple 'threads'. Certain processes in EasyDIVER (such as joining with PANDAseq) can utilize this capability to run faster. As a rule of thumb, you may set the number of threads equal to the number of cores on your machine for optimal performance. For further optimization, look up the specifications of your hardware and adjust your thread count accordingly.

    - **Example:** In the example command provided in this manual, the thread count is set as `-T 14`, which would be a desirable choice for a machine with 16 threads. To be safe, and assuming your machine is at least a quad-core CPU, use `-T 4`.