

---

# **shear Documentation**

***Release 2017-06-18***

**James B. Pease**

**Jun 20, 2017**



**CONTENTS:**

<b>1</b>	<b>Getting Started</b>	<b>1</b>
1.1	What is SHEAR? . . . . .	1
1.2	Requirements . . . . .	1
1.3	Installation . . . . .	1
1.4	Preparing your data . . . . .	1
1.5	Usage . . . . .	2
<b>2</b>	<b>Program Parameter Descriptions</b>	<b>3</b>
2.1	adapt . . . . .	3
2.2	shear . . . . .	4
<b>3</b>	<b>Indices and tables</b>	<b>11</b>



## GETTING STARTED

### 1.1 What is SHEAR?

SHEAR (Simply Handler for Error and Adapter Removal) is a short-read trimmer for high-throughput sequencing fastq files. SHEAR first scans the fastq file(s) and automatically detects likely adapter and primer contaminants. Then it calls Scythe (<https://github.com/vsbuffalo/scythe>), which removes reads using a Bayesian error-tolerant approach that effectively removes adapters even if the adapter itself contains a sequence variation due to sequencing error. SHEAR then trims and filters reads based on minimum quality and content cutoffs. Additionally, SHEAR is designed to automate the simultaneous trimming and concatenation of multiple paired read files into a single trimmed file ready for mapping or assembly.

### 1.2 Requirements

- Python 2.7.x or 3.x (3.x recommended)

#### 1.2.1 Optional

- Scythe: <https://github.com/vsbuffalo/scythe> (**Strongly recommended**)

### 1.3 Installation

No installation is necessary, simply clone the repository from GitHub. Scythe should be installed according to its instructions.:

```
git clone https://www.github.com/jbpease/shear
```

### 1.4 Preparing your data

Standard fastq files with four lines per entry (header, sequence, gap, quality) should be used. ABI solid colorspace reads are not currently supported. When using paired-end mode reads only need to be sorted and contain the same number of reads if the `-U/--filter-unpaired` mode is used, since this will remove both reads from a pair when either of them is filtered out.

## 1.5 Usage

### 1.5.1 Paired-end

```
python shear.py --fq1 FASTQ.SAMPLE1.p1.fastq FASTQ.SAMPLE2.p1.fastq ... --  
↪fq2 FASTQ.SAMPLE1.p2.fastq FASTQ.SAMPLE2.p2.fastq ... --out1 FASTQ.  
↪sheared.p1.fq --out2 FASTQ.sheared.p2.fq
```

### 1.5.2 Single-read

```
python shear.py --fq1 FASTQ.SAMPLE1.p1.fastq FASTQ.SAMPLE2.p1.fastq ... --  
↪fq2 FASTQ.SAMPLE1.p2.fastq FASTQ.SAMPLE2.p2.fastq ... --out1 FASTQ.  
↪sheared.p1.fq --out2 FASTQ.sheared.p2.fq
```

### 1.5.3 Config file alternative

Alternatively to a full set of command line arguments you can enter a single positional argument that points in a text file. You can then specify command line arguments more neatly over several lines as in the example:

```
--fq1  
FASTQ.SAMPLE1.p1.fastq  
FASTQ.SAMPLE2.p1.fastq  
--fq2  
FASTQ.SAMPLE1.p2.fastq  
FASTQ.SAMPLE2.p2.fastq  
--out1 FASTQ.sheared.p1.fq  
--out2 FASTQ.sheared.p2.fq
```

## PROGRAM PARAMETER DESCRIPTIONS

### 2.1 adapt

#### 2.1.1 Description

This program (as part of SHEAR) searches adapter sequences and generates an adapter file for use with SHEAR/Scythe.

#### 2.1.2 Parameters

**-h/--help**

**Description:** show this help message and exit

**Type:** boolean flag

**--fq1 (required)**

**Description:** one or more fastq file paths, separated by spaces

**Type:** file path; **Default:** None

**-o/--out (required)**

**Description:** output FASTA of adapters detected

**Type:** file path; **Default:** None

**--fq2**

**Description:** one or more fastq file paths separated by spaces, only use this for paired-end fastq files and enter these files in the same order as their counterparts in -fq1

**Type:** file path; **Default:** None

**-k/--end-klength**

**Description:** Length of end kmer to tabulate for possible adapter matches.

**Type:** integer; **Default:** 16

**-m/--mode**

**Description:** known=only use list of known adapters;endmer=search for common 3' end sequences;both=both known and endmers

**Type:** None; **Default:** known

**Choices:** ('known', 'endmer', 'both')

**--quiet**

**Description:** Suppress progress messages

**Type:** boolean flag

**-M/--min-match**

**Description:** Minimum proportion of read match required to report the kmer as a possible match.

**Type:** integer; **Default:** 0.0001

**-N/--number-of-reads**

**Description:** Number of reads to search in each fastq

**Type:** integer; **Default:** 200000

## 2.2 shear

### 2.2.1 Description

SHEAR is a read trimmer that coordinates the automatic finding of adapter sequences, removes adapters using Scythe, implements various trimming and filtering options for high-throughput short read sequences, and allows coordinated removal of paired end sequence files.

### 2.2.2 Parameters

**-h/--help**

**Description:** show this help message and exit

**Type:** boolean flag

**--fq1 (required)**

**Description:** one or more fastq file paths, separated by spaces

**Type:** file path; **Default:** None



**--out1 (required)**

**Description:** Output fastq file path. Note this is a single output file that concatenates the processed outputs from all files in `-fq1`

**Type:** file path; **Default:** None

**-a/--adapters**

**Description:** Skip adapter finding and use these adapter files. Either enter (1) a single file to use for all fastq files, (2) one file per single end file or pair of paired-end files. Adapter file(s) should be in FASTA format.

**Type:** file path; **Default:** None

**--clean-header**

**Description:** removes any additional terms from the header line (useful after STAR)

**Type:** boolean flag

**-f/--trim-fixed**

**Description:** Trim a fixed number of bases from the FRONT:END of the sequence (NOT recommended).

**Type:** None; **Default:** 0:0

**--filt1**

**Description:** Output fastq file path for sequences that were filtered out. Note this is a single output file that concatenates the rejected outputs from all files in `-fq1`. Default is `[-out1]_filtered_1.fastq`

**Type:** file path; **Default:** None

**--filt2**

**Description:** Output fastq file path for sequences that were filtered out. Note this is a single output file that concatenates the rejected outputs from all files in `-fq2`. Default is `[-out2]_filtered_2.fastq`

**Type:** file path; **Default:** None

**--fq2**

**Description:** one or more fastq file paths separated by spaces, only use this for paired-end fastq files and enter these files in the same order as their counterparts in `-fq1`

**Type:** file path; **Default:** None

**-k/--adapter-end-klength**

**Description:** (Adapter finding) Length of end kmer to tabulate for possible adapter matches.

**Type:** integer; **Default:** 16

### `--log-path`

**Description:** Manually specify log file path, default is 'shear\_TIMESTAMP'

**Type:** file path; **Default:** None

### `-m/--adapter-mode`

**Description:** (Adapter finding) known=only use list of known adapters;endmer=search for common 3'end sequences;both=both known and endmers

**Type:** None; **Default:** known

**Choices:** ('known', 'endmer', 'both')

### `-n/--retain-ambig`

**Description:** By default ambiguous nucleotides (N) are removed from both ends of each read. If this flag is specified, N's are retained.

**Type:** boolean flag

### `--out2`

**Description:** output fastq file path. Note this is a single output file that concatenates the processed outputs from all files in -fq2

**Type:** file path; **Default:** None

### `-p/--trim-poly`

**Description:** Trim poly-A or poly-T repeats of at least this length from the front or end.

**Type:** integer; **Default:** 12

### `-q/--trim-qual`

**Description:** Trim bases below this quality score from the FRONT:END of each read.

**Type:** None; **Default:** 20:20

### `--quality-scale`

**Description:** Quality scale is usually automatically determined, but use this to set manually.

**Type:** None; **Default:** None

**Choices:** ('sanger', 'illumina', 'phred', 'solexa')

### `--quiet`

**Description:** Suppress progress messages

**Type:** boolean flag

**--retain-temp**

**Description:** Retain temporary files (none=remove all; tempfastq=remove temporary fastq files from scythe; exceptadapters=remove temp fastq and log files from Scythe but keep adapter file; all=keep all temporary file)

**Type:** None; **Default:** none

**Choices:** ['none', 'tempfastq', 'exceptadapters', 'all']

**-s/--scythe-prior**

**Description:** Bayesian prior for proportion of adapters expected to be sampled in Scythe.

**Type:** float; **Default:** 0.1

**--scythe-match**

**Description:** Minimum number of bases required for a match in Scythe.

**Type:** integer; **Default:** 5

**--scythe-skip**

**Description:** Skip scythe 3' adapter removal.

**Type:** boolean flag

**-t/--platform**

**Description:** Sequencing Platform

**Type:** None; **Default:** TruSeq

**Choices:** ('TruSeq', 'TruSeqDualIndex')

**--temp-dir**

**Description:** directory to use for temporary files

**Type:** file path; **Default:** .

**--trim-qual-pad**

**Description:** Trim additional bases next to low-quality bases specified by -trim-qual from the FRONT:END of each read.

**Type:** None; **Default:** 0:0

**-y/--trim-pattern-5**

**Description:** Comma-separated list of specific sequences to trim from the 5' end. (Not recommended).

**Type:** None; **Default:** None

**-z/--trim-pattern-3**

**Description:** Comma-separated list of specific sequences to trim from the 3' end. Can be used for extra stringent adapter trimming.

**Type:** None; **Default:** None

**-A/--filter-ambig**

**Description:** Filter reads with more than this number of ambiguous nucleotides (N's; set as 0 to skip

**Type:** integer; **Default:** 5

**-I/--filter-low-info**

**Description:** Filter out reads with mutual information scores exceeding this value (ADVANCED, removes highly repetitive reads).

**Type:** float; **Default:** 0.0

**-L/--filter-length**

**Description:** Filter out reads that contain fewer than this many characters after trimming.

**Type:** integer; **Default:** 30

**-M/--adapter-min-match**

**Description:** (Adapter finding) Minimum proportion of read match required to report the kmer as a possible match.

**Type:** integer; **Default:** 0.0001

**-N/--adapter-number-of-reads**

**Description:** (Adapter finding) Number of reads to search in each fastq

**Type:** integer; **Default:** 200000

**-Q/--filter-quality**

**Description:** Filter out reads with a mean quality score below this value (before trimming).

**Type:** integer; **Default:** 3

**-U/--filter-unpaired**

**Description:** If either read in a read pair is filtered out, the counterpart reads is also filtered out regardless of quality.

**Type:** boolean flag

**-X/--scythe-executable**

**Description:** Set the path of the scythe executable manually.

**Type:** None; **Default:** scythe



## INDICES AND TABLES

- `genindex`
- `modindex`
- `search`