Fastq\_clean : Manual

*Fei lab*

*Boyce Thompson Institute for Plant Research,*

*Cornell University,*

*Tower Road, Ithaca, NY 14853,*

*United States of America,*

Oct 12, 2014

[1. Introduction 2](#_Toc402892392)

[2. Download and Installation 2](#_Toc402892393)

[2.1 Software Requirements 2](#_Toc402892394)

[2.2 Parameters Related to Hardware 2](#_Toc402892395)

[2.3 Begin with An Example 3](#_Toc402892396)

[3. Processing Illumina Sequencing data 4](#_Toc402892397)

[3.1 Principal 4](#_Toc402892398)

[3.2 The Pipeline 5](#_Toc402892399)

[3.3 Parameters 7](#_Toc402892400)

[3.3.1 Required Parameters 7](#_Toc402892401)

[3.3.2 Optional Parameters 8](#_Toc402892402)

[3.3.3 R-script related options 9](#_Toc402892403)

[3.3.4 BWA related options 10](#_Toc402892404)

[3.3.5 Tweaking with different read length 10](#_Toc402892405)

[4. Quality Control 11](#_Toc402892406)

[4.1 Quality Reporting 11](#_Toc402892407)

[4.2 Quality Analysis 12](#_Toc402892408)

[5. Runtime errors and solutions 13](#_Toc402892409)

[6. Reference 14](#_Toc402892410)

**Notice：**

Fastq\_clean Toolkit is available at [appbox](https://app.box.com/s/v7sjjmjo56r71k8bvqkv) or [baiduPan](http://pan.baidu.com/s/1sjllrcD). You can download and use it only for academic purposes with a requirement to cite our published paper[1]. Any commercial use or distribution is forbidden. For any questions or error reports, please contact [Chengji chen](mailto:chengjiechen@stu.scau.edu.cn).

## Introduction

Fastq\_clean is an optimized pipeline to clean the DNA-seq and RNA-seq data from the  
illumina sequencer. It is implemented to remove the low quality nucleotides and adapters precisely and keep as many of the qualified nucleotides as possible. Support for data from other sequencers(e.g. ion torrent) was added to Fastq\_clean version 2.0.

Fastq\_clean can deal with paired-end sequencing data provided by Illumina sequencer,which is the sole equipment that provide paired-end data in practice.

## Download and Installation

### 2.1 Software Requirements

1. Linux Platform only for now

2.Perl 5.0+ is required

3. R/Bioconductor is required

Other Prerequisite tools such as BWA and Samtools was already included.

You can install R by entering the commands:

|  |
| --- |
| #for Fedora,Centos or RHEL  $sudo yum install R-devel  #for Debian,Ubuntu  $sudo apt-get install r-base-dev |

Install Bioconductor by entering

|  |
| --- |
| $R  $source("http://bioconductor.org/biocLite.R");  $biocLite();  $biocLite("ShortRead"); |

### 2.2 Parameters Related to Hardware

Fastq\_clean is implemented to meet minimal Hardware Requirements. User can tuning performance for specific hardware by setting two parameters. Parameter "thread\_num" refers to the number of worker threads and parameter "read\_PerYield" set the Number of reads processed in each iteration to limit the memory usage. So the total memory used by Fastq\_clean can be calculated as:

read\_PerYield \* 4 \* 100

For example,if "read\_PerYield" is set to 5M,you need at least 5M\*4\*100=2G RAM to run the program.

The "thread\_num" and the "read\_PerYield" is set to 8 and 5M by default.

Notice that Data processing could slow down dramatically by a small " read\_PerYield ".

### 2.3 Begin with An Example

This section will guide you to clean the chrysanthemum sequence data with Fastq\_clean step by step. The dataset contains paired-end sequencing data from illumina sequencer. All the six data files can be retrieve from NCBI SRA database.

1. Download Fastq\_clean package and extract it, rename the directory as "fastq\_clean"

|  |
| --- |
| $ mv fastq\_clean-X.Y fastq\_clean |

2. Change to the directory and make all the scripts in it executable

|  |
| --- |
| $ cd fastq\_clean  $ chmod +x -R . |

3. Decompress the database files in the sub directory databases

|  |
| --- |
| $ gunzip ./databases/\*.gz |

4. Download all the six data files from the NCBI SRA database into the working directory and transform them into fastq format:

|  |
| --- |
| $ wget ftp-trace.ncbi.nlm.nih.gov/sra/sra-instant/reads/ByRun/sra/SRR/SRR921/SRR921340/SRR921340.sra  $ wget ftp-trace.ncbi.nlm.nih.gov/sra/sra-instant/reads/ByRun/sra/SRR/SRR921/SRR921341/SRR921341.sra  $ wget ftp-trace.ncbi.nlm.nih.gov/sra/sra-instant/reads/ByRun/sra/SRR/SRR921/SRR921342/SRR921342.sra  $ wget ftp-trace.ncbi.nlm.nih.gov/sra/sra-instant/reads/ByRun/sra/SRR/SRR921/SRR921321/SRR921321.sra  $ wget ftp-trace.ncbi.nlm.nih.gov/sra/sra-instant/reads/ByRun/sra/SRR/SRR921/SRR921322/SRR921322.sra  $ wget ftp-trace.ncbi.nlm.nih.gov/sra/sra-instant/reads/ByRun/sra/SRR/SRR921/SRR921324/SRR921324.sra  # transform .sra to .fastq  $ nohup ./tools/fastq-dump SRR921340.sra  $ nohup ./tools/fastq-dump SRR921341.sra  $ nohup ./tools/fastq-dump SRR921342.sra  $ nohup ./tools/fastq-dump SRR921321.sra  $nohup ./tools/fastq-dump SRR921322.sra  $nohup ./tools/fastq-dump SRR921324.sra |

5. Run the pipeline to clean the six data files:

|  |
| --- |
| nohup ./fastq\_clean.pl --file\_list fastqfiles1.list --run\_R --pattern\_file patterns\_forward --rRNA\_removal --rRNA\_reference rRNA\_silva111.fasta --virus\_removal --virus\_reference vrl\_genbank.fasta |

## Processing Illumina Sequencing data

### 3.1 Principal



fig.1 Illumina Adapter Sequence

Illumina sequencer separates the sequenced read into three regions: the 5’ low quality region, the high quality region located in the middle of the read, and the 3’ low quality region (see fig.1).

Multiplex technology was introduced to process multiple samples in 1 Lane simultaneous.More specifically, each sample was given a unique Barcode which is a 6bp or 8bp DNA sequence.The barcode can used to recognized different samples.

In fig.1,S1 is the forward sequencing primer, the residue next to it located at the beginning of the forward sequence; S2 is the reverse sequencing primer,the the residue next to it located at the beginning of the reverse sequence.

Single end sequencing means to sequence data from "S1" Side. Paired-end sequencing means to sequence from both "S1" and "S2" sides. The Barcode was obtained by sequencing "Sb" primer individually.

Theoretically, peak value of RNA-seq library insert length is 200bp or 300bp,so the sequencing should get some piece of target sequence form 5' end.However, for those target sequences shorter than Read length(e.g. ,100bp), sequencing process will match the 3' end adapter improperly, so called adapter con contamination.

During pre-processing steps,if too much adapter sequence was found typically due to the inappropriate RNA-seq library insert length.And if there are many full-length 3' adapter sequences,it usually comes from self ligation. So use TruSeq RNA Sample Preparation Kits,we can find the information as follow.

|  |
| --- |
| PCR primer 1  AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT  PCR primer 2 reverse complement  AGATCGGAAGAGCACACGTCTGAACTCCAGTCACNNNNNNATCTCGTATGCCGTCTTCTGCTTG  PCR primer 2  CAAGCAGAAGACGGCATACGAGATNNNNNNGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT  PCR primer 1 reverse complement  AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTAGATCTCGGTGGTCGCCGTATCATT  Target sequence in the forward sequencing fastq file  AGATCGGAAGAGCACACGTCTGAACTCCAGTCACNNNNNNATCTCGTATGCCGTCTTCTGCTTG  Target sequence in the reverse sequencing fastq file  AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTAGATCTCGGTGGTCGCCGTATCATT |

### 3.2 The Pipeline

Fastq\_clean toolkit provide a set of scripts (in the sub directory "tools") to generate directories and files by the TruSeq adapter sequence structure automatically. The processes are describe as follow.

1. Copy or move all the data files for processing to $work\_dir，make sure that all the sequence data files have the suffix ".fastq",and the file names of forward sequencing data file should end with "\_1"(e.g., "sample1\_1.fastq) while the reverse sequencing data file names should end with "\_2(e.g.,"sample\_2.fastq").

2.create a list file contains names of all the data file.

|  |
| --- |
| cd fastq\_clean  ./tools/getFileNames.pl >samples.list  #manual checking of the list file is highly recommend here  cat samples.list |

3. Extracting the barcode list（the first column of barcode list is the sample's name,the second column is the corresponding barcode).Because the auto-generated barcode is recognized by the first read in sequence data ,there may be some sequencing errors(the accuracy could be improved by searching all the reads instead of the first one).So the barcode list must be contrast manually with the barcode provided by sequencing sevice company.

|  |
| --- |
| ./tools/getBarcodes.pl --file\_listsamples.list>barcodes.list  cat barcodes.list |

if the barcode can not be extract automatically，it can be added by hand.

|  |
| --- |
| sample1\_1 barcode1  sample2\_1 barcode2  sample1\_1 barcode1  sample2\_1 barcode2 |

4.confirm the adapter /primer sequence,take data from TruSeq kits as an example

|  |
| --- |
| #forward 3' end adapter for all the sequence files,the forward sequencing result should not contais zero while all the reverse sequencing result should be all zero  ./tools/checkAdapters.pl --file\_listbarcodes.list --adapter AGATCGGAAGAGCACACGTCTGAACTCCAGTCAC  #forward 3' end adapter for all the sequence files,the reverse sequencing result should not contais zero while all the forward sequencing result should be all zero  ./tools/checkAdapters.pl --file\_listbarcodes.list --adapter AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTA |

It is recommend to check a forward sample and a reverse sample manually(optional step).

|  |
| --- |
| # check a forward sequencing data file sample1\_F.fastq  grep AGATCGGAAGAGCACACGTCTGAACTCCAGTCAC sample1\_F.fastq> 1  wc -l 1  # check a reverse sequencing data file sample1\_R.fastq  grep AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTA sample1\_R.fastq> 2  wc -l 2 |

5. the forward and revese sequencing results in two individual files

|  |
| --- |
| ./tools/split\_filelist.pl --in barcodes.list --out1 barcodes1.list --out2 barcodes2.list |

6. create the fastqfiles1.list

|  |
| --- |
| ./tools/concatAdapters.pl --file\_list barcodes1.list --prefix AGATCGGAAGAGCACACGTCTGAACTCCAGTCAC --suffix ATCTCGTATGCCGTCTTCTGCTTG --with\_barcode> fastqfiles1.list  cat fastqfiles1.list |

create the fastqfiles1.list

|  |
| --- |
| ./tools/concatAdapters.pl --file\_list barcodes2.list --prefix AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTAGATCTCGGTGGTCGCCGTATCATT > fastqfiles2.list  cat fastqfiles2.list |

7. processing the forward end sequencing data

|  |
| --- |
| nohup ./illumina\_clean.pl forward.config |

8.processing the reverse end sequencing data

|  |
| --- |
| nohup ./illumina\_clean.pl reverse.config |

9. get the match reads (optional step)

Because the paired-end sequencing produces a lot of mismatch reads,we can use the script "match\_paired.pl" to get matched pair data and write the mismatch ones to another file.

|  |
| --- |
| ./tools/match\_paired.pl sample1\_1sample1\_2 |

The results consist of a match file (e.g.,"sample1\_1.fq) and a mismatch file (e.g.,"sample1\_S.fq).

### 3.3 Parameters

There are four categories of parameters which can be used to customize the pipeline of Fastq\_clean : Required,Optional,R language related and BWA related.

#### 3.3.1 Required Parameters

--file\_list

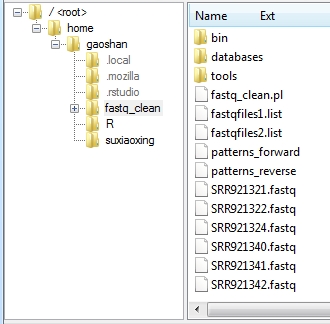
The fastq\_clean program is designed to process input files in batch, so a text file should be provided to include a list of input file names (the first column) and the corresponding 3’ adapter sequences (the second column).



**fig.2 The constant of fastqfile1.list**

Two input list files, “fastqfiles1.list” and “fastqfiles2.list” are included in the Fastq\_clean package for demonstration of the pipeline(see section 2.3).

Take "fastqfiles1.list" as an example(see fig. 2), the first column in the file are the input files' names without suffix,and the second column are 3’ adapter sequences provided by TruSeq RNA Sample Preparation Kits, which covers more than 80% usage in RNA-seq and also used in some DNA-seq application. Users can simply replace the first column with their own data file names and replace the second column with their own adapter sequences to run this pipeline on their own data.



**Fig. 3. The arrangement of the working directory**

The corresponding input files in the list must be included in the working directory with suffix name “.fastq”(see fig.3). All the input files must be the raw data from the sequencer. The raw data having lines of the same length should not be processed by any data clean programs.

#### 3.3.2 Optional Parameters

-- run\_R

run the R script to remove low quality, ambiguous and adapter bases in the input data. The default setting is to run the R script.

-- rRNA\_removal

run the Perl script to remove reads which can be aligned to the rRNA reference sequences.

Notic that all the data file must have a suffix name ".clean"

-- rRNA\_reference

If the parameter “rRNA\_removal” is set, an rRNA reference database named rRNA\_silva111.fasta in FASTA format should be provided in the folder $work\_dir/databases.

-- virus\_removal

To run the Perl script to remove reads which can be aligned to the virus reference sequences.

-- virus \_reference

To run the Perl script to remove reads which can be aligned to the virus reference sequences.

#### 3.3.3 R-script related options

--pattern\_file

The name of a file containing a list of adapter pattern strings. There are two pattern files “patterns\_forward” and “patterns\_reverse” in the working directory. They are designed to remove some other contamination in the forward and reverse sequencing data after removing the low quality, ambiguous and adapter bases. These patterns are also from TruSeq RNA Sample Preparation Kits.

--quality\_Cutoff

Low quality bases below this parameter will be trimmed from both ends of reads. The default value is 20, which is the most common used in quality control.

--region\_5Len

The low quality region from the 5' end of reads. The default value is10. The “region\_5Len”, “region\_3Len”, and “n\_Cutoff” are optimized for the illumine 100bp sequencing technology.

--region\_3Len

The low quality region from the 3' end of reads. The default value is 20.

--n\_Cutoff

After low quality bases trimmed, the reads containing "N" with a number above this parameter will be removed. The default value is 2.

--adapter\_mismatch

A mismatch ratio between adapters with the 3' ends of reads. The default value is 0.1.

--read\_Length

Reads with the length shorter than this parameter will be removed. The default value is 25.

--read\_PerYield

Number of reads processed in each iteration to limit the memory usage. The default value is 5e5.

#### 3.3.4 BWA related options

--max\_dist

Maximum edit distance,default value is 2

--max\_open

Maximum number of gap opens ,default value is 1

--max\_extension

Maximum number of gap extensions,default value is 1

--len\_seed

Take the first INT subsequence as seed,default value is 50

--dist\_seed

Maximum edit distance in the seed,default value is 1

--thread\_num

Number of threads (multi-threading mode) ,default value is 8

This pipeline uses the BWA program [[2](#_ENREF_2)] to align reads to the virus or rRNA reference sequences. To call the BWA program in the command line, six parameters (“n”, “o”, “e”, “l”, “k”, “t”) need be set and renamed by (“max\_dist”, “max\_open”, “max\_extension”, “len\_seed”, “dist\_seed”, “thread\_num”). For detailed description of those parameters, users can see the online manual of the program BWA (http://bio-bwa.sourceforge.net/bwa.shtml).

#### 3.3.5 Tweaking with different read length

The “region\_5Len”, “region\_3Len”, and “n\_Cutoff” parameters (see 3.3.2) are optimized for the illumine 100bp sequencing technology. For seq data of other length, users must set these parameters by themselves. Table.1 and table.2 provides some recommended parameters for user’s convenience.

Table. 1 parameters for seq data of different length

|  |  |  |  |
| --- | --- | --- | --- |
| Length(bp) | region\_5Len | region\_3Len | n\_Cutoff |
| <=50 | 10 | 10 | 1 |
| 50-100 | 10 | 20 | 2 |
| 150-200 | 10 | 50 | 3 |
| >=250 | 10 | 70 | 4 |

Table. 2 BWA related parameters on case of different read length

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Length(bp) | max\_dist | max\_open | max\_extension | len\_seed | dist\_seed |
| <=50 | \*2/1 | 1 | 1 | 35 | 1/1 |
| 50-100 | 4/2 | 1 | 1 | 50 | 2/1 |
| 150-200 | 6/3 | 2 | 1 | 70 | 3/2 |
| >=250 | 8/4 | 2 | 1 | 100 | 4/2 |

\*Parameters on the left apply to application required high quality results such as De novo assembly. The Parameters on the right can be used for application not sensitive to the dat accuracy such as aligning .

## Quality Control

### 4.1 Quality Reporting

The most important feature of Fastq\_clean is the statistics reporting（see fig.4 and fig.5) .it provided with the cleaning . These comprehensive information can help users understand what happened in the pipeline and track the problems.

Fig.4and Fig.5 show the Quality Control Information Report. Columns 1 to 7 in the report come from file "trimmed.report".Columns 8 to 9 are extracted from fil "clean.report". Column 14 can be obtained by running the script "match\_paired.pl". User can copy all the information into a MS excel template (in the doc sub directory) for convenience.



fig. 4 Quality Control Information Report（column 1 to 11)

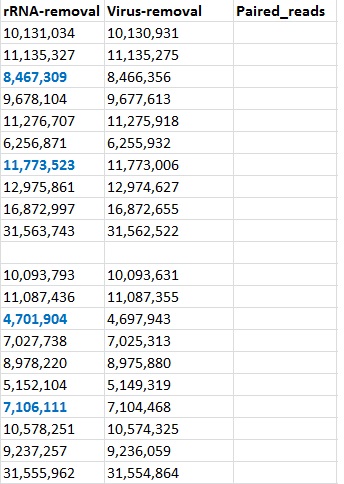


fig 5. Quality Control Information Report（column 12 to 14)

The explanation of each column (1 to 14) of the quality report is describe as below.

Sample：file name of samples without suffix；

N\_number\_in\_raw\_data：counts of undetermined bases in the raw data;

Raw\_reads: reads counts of the raw data;

Raw\_len：read length of the raw data；

High\_quality\_reads：High quality reads(Trimming the low quality nucleotides and Ns);

Trimmed\_reads：counts of high quality reads that longer than parameter "readLength" after removing 3' end adapt contamination;

Trimmed\_len：average lenth of Trimmed\_reads;

Clean\_reads：Trimmed reads left after removing those which match patterns (the patterns comes from files "patterns\_forward" and "patterns\_reverse");

Clean\_length：average lenth of Clean\_reads;

Removed\_reads：ratio of the removed reads by rules described above,calculated by(clean\_reads-Raw\_reads)/Raw\_reads；

Removed\_nt: ratio of the removed bases by rules described above, calculated by（Clean\_reads\*Clean\_length-Raw\_reads\*Raw\_len）/Raw\_reads\*Raw\_len;

rRNA-removal：reads left after removing rRNA sequences;

Virus-removal：reads left after removing Virus sequences;

Paired-reads：matched reads of forward sequences and reverse sequences in paired-end sequencing;

### 4.2 Quality Analysis

Columns of the Quality Control Report allow us to analysis the quality of data in a intuitive way. Generally speaking,column "Removed\_reads" is less than 5% and column "Removed\_nt" is less than 10% show that the raw data are in high quality. If "Removed\_reads" is less than 10% and "Removed\_nt" is less than 20% ,the quality of raw data are acceptable.

Data Changes form a column to the next one in Quality Control Report give rich information to analysis the quality.

1. IF column "Raw\_reads" to column "High\_quality\_reads" decrease significantly（rarely happens on our experience）, the accuracy of data sequencing is doubtful,especially the high quality reads adjacent to 5' end adapter.

2.IF column "High\_quality\_reads" to column "Trimmed\_reads" decrease significantly,it may infer too small insert size of sequence library or abnormal RNA-seq degradation (in rare case).Notice that too small read size can't be assembly efficiently.

3. IF column "Trimmed\_reads" to column "Clean\_reads" decrease significantly,it usually means self-ligation by insufficient RNA samples.

4.IF column "Clean\_reads" to column "rRNA-removal" decrease significantly,it reflect severe rRNA Contamination which will sacrifice the RNA-seq quality.

5. IF column "Clean\_reads" to column "Virus-removal" decrease significantly,it usually means severe virus Contamination and the data was useless（doesn't happen very often on our experience）.

## Runtime errors and solutions

1. "Input file having line of different length"

|  |
| --- |
| Error in replaceLetterAt(seqs, at, letter) :  'x' must be rectangular (i.e. have a constant width) |

**Cause：**The raw data must contain reads of the same length. Some sequence service provider cleaned data before they sent you.

**Solution**: We strongly suggest you ask the raw data not only for using this pipeline, but also for publication use.

2.Rscript gives out error code

**Cause:** bug exists in the newest R(version 3.1.1)

**Solution:**try the R code as below：

|  |
| --- |
| a<-readFastq("sample1\_1.fastq")  writeFastq(a, file=" sample1\_1.fq ", mode="w", full=FALSE);  #content of sample1\_1.fastq  @FCC3215ACXX:7:1101:1378:2068#/1  TGCGGCGAAGTCGCGGATCTTTCGACCTCCTCGAGCGCCTCCTCCCGCCCAAGACTTCGGCGAGAGAGGGACAGTGGAGTACGAGCTCGA  +  aabacceeggceghhiiiiiiiiiiiihggggedeccccZ\_abbccaaccWacccccbccac[a]aacccZ\_^aRY^bcJX`aaac\_baa |

## Reference

1. Xu, Y., et al., Transcriptome sequencing and whole genome expression profiling of chrysanthemum under dehydration stress.BMC Genomics, 2013.14(1): p. 662.

2. Li, H. and R. Durbin, Fast and accurate short read alignment with Burrows-Wheeler transform.Bioinformatics, 2009.**25**(14): p. 1754-1760.