The Journal of Microbiology (2012) Vol. 50, No. 5, pp. 766–769
DOI 10.1007/s12275-012-2494-6

PyroTrimmer: a Software with GUI for Pre-Processing 454 Amplicon Sequences

Jeongsu Oh[1,2], Byung Kwon Kim[3], Wan-Sup Cho[2], Soon Gyu Hong[4], and Kyung Mo Kim[1]*

[1]*Biological Resource Center, Korea Research Institute of Bioscience and Biotechnology, Daejeon 305-806, Republic of Korea*
[2]*Department of Bio-Information Technology, Chungbuk National University, CheongJu 361-763, Republic of Korea*
[3]*Department of Systems Biology, Yonsei University, Seoul 120-749, Republic of Korea*
[4]*Division of Polar Life Sciences, Korea Polar Research Institute, Incheon 406-840, Republic of Korea*

(Received September 10, 2012 / Accepted September 26, 2012)

The ultimate goal of metagenome research projects is to understand the ecological roles and physiological functions of the microbial communities in a given natural environment. The 454 pyrosequencing platform produces the longest reads among the most widely used next generation sequencing platforms. Since the relatively longer reads of the 454 platform provide more information for identification of microbial sequences, this platform is dedicated to microbial community and population studies. In order to accurately perform the downstream analysis of the 454 multiplex datasets, it is necessary to remove artificially designed sequences located at either ends of individual reads and to correct low-quality sequences. We have developed a program called PyroTrimmer that removes the barcodes, linkers, and primers, trims sequence regions with low quality scores, and filters out low-quality sequence reads. Although these functions have previously been implemented in other programs as well, PyroTrimmer has novelty in terms of the following features: i) more sensitive primer detection using Levenstein distance and global pairwise alignment, ii) the first stand-alone software with a graphic user interface, and iii) various options for trimming and filtering out the low-quality sequence reads. PyroTrimmer, written in JAVA, is compatible with multiple operating systems and can be downloaded free at http://pyrotrimmer.kobic.re.kr.

Keywords: 454, pre-processing, pyrosequencing, trimming, software

Introduction

Traditional microbial ecology relies on culture-based methods. However, the majority of microorganisms cannot be cultured in laboratories (Pace *et al.*, 1985; Xu, 2006). Metagenomics, a term first coined in 1998, is the study of genetic material extracted directly from natural environments (Handelsman *et al.*, 1998). This approach enables us to study directly both the culturable and unculturable microbes in their native environment. In the past, many metagenome projects have utilized Sanger sequencing and cloning. However, these methods cannot produce the large number of reads because of their cost and technical limitations. With the recent advent of next generation sequencing (NGS) technology, high-throughput sequencing of metagenomes at reasonable cost has become possible.

The 454 pyrosequencing platform provides longer reads (e.g., recently over 600 bp in length) than other popular next generation sequencing (NGS) platforms such as Illumina and Solid. Since longer reads generally increase the reliability of sequence comparisons, they are considered superior for species identification as well as gene annotation of metagenomic sequences. For this reason, the 454 platform has been widely used in metagenomic approaches to studying microbial community structures. This approach usually involves PCR of the target gene of multiple samples and thus necessitates the need to remove barcodes, linkers, and primers that are incorporated during the multiplex amplification. In addition, ambiguous nucleotides, low quality reads and short reads are also problematic and need to be trimmed or filtered out.

Pre-processing includes trimming the tag sequences that are added during the PCR amplification and the filtering out of the low-quality reads. Many tools for this purpose have been developed previously such as CANGS (Pandey *et al.*, 2010), PyroTagger (Kunin and Hugenholtz, 2010), RDP (Cole *et al.*, 2009), and TagCleaner (Schmieder *et al.*, 2010). Since these tools are mostly part of large pipelines, they are not dedicated applications for pre-processing the 454 reads. Therefore, they lack effective functionalities for trimming and filtering out low-quality sequence reads with high sensitivity. Moreover, these tools are command line or web applications that are fairly complicated for common use. The command line applications are platform-dependent and sometimes require installations of additional programs for proper functioning. In addition, web systems take a longer time to upload the high-throughput 454 data. Moreover, the absence of a graphic user interface (GUI) discourages an intuitive user-approach to the application. To overcome these limitations, we developed PyroTrimmer, a user-friendly software with a GUI for pre-processing multiplex pyrosequencing datasets. PyroTrimmer supports various functions for trimming and filtering out erroneous reads accurately and efficiently. Having 454 reads pre-processed by PyroTrimmer makes the downstream analysis more reliable.

*For correspondence. E-mail: kmkim@kribb.re.kr; Tel.: +82-42-860-4613; Fax: +82-42-860-4625

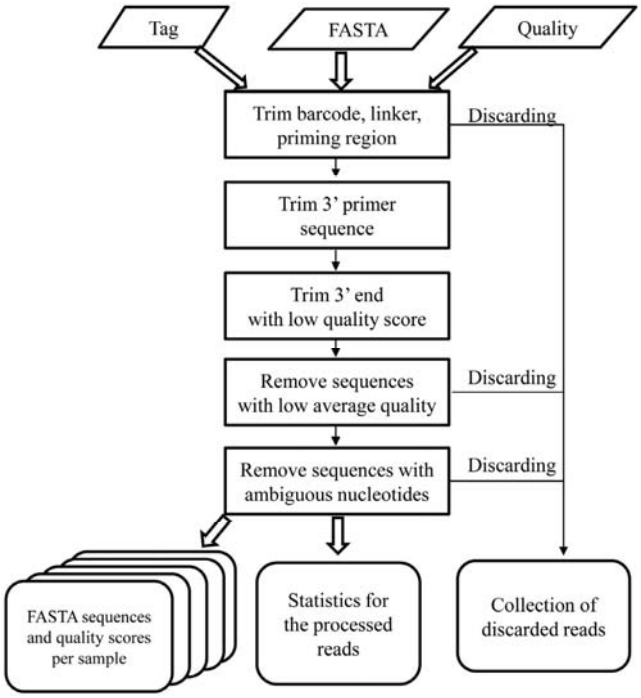

Fig. 1. Schematic representation of PyroTrimmer. The diagram illustrates major tasks, as well as input and output data types.

Materials and Methods

Input

PyroTrimmer takes a FASTA sequence file, quality file and tag information as input (Fig. 1). A FASTA file contains 454 pyrosequencing raw reads, each of which usually has tags such as barcode, linker, and primer sequences at the 5′ end. If sequence reads are long enough to cover the entire amplicons, tags can also be present in the 3′ end of the reads. A quality file keeps quality scores for individual bases of the corresponding sequences in the FASTA file. A tag information file includes names of projects, samples and genes, upper and lower bounds of read length cutoffs, and sequences of barcodes, linkers and primers. More detailed information can be found in the user guide, which is available at http://pyrotrimmer.kobic.re.kr.

Trimming barcode, linker and primer

Since the 454 platform supports multiplexing of a number of samples (Thomas *et al.*, 2012), multiple barcodes are used to identify reads per sample in a run of the machine. PyroTrimmer can manipulate the multiplex dataset regardless of sequencing direction. This program examines the presence of barcodes, linkers, and primers in a given FASTA sequence file by referring to the input tag information. It further identifies whether barcodes and linkers provided by the tag file are present in the 5′ end of reads by using exact string match (Fig. 1). The primer sequences are then compared with the 5′ regions of reads. Since the 454 reads can have errors from nucleotide substitution, PyroTrimmer allows nucleotide mismatches that are below the given cut-off option through the Needleman-Wunsch algorithm (Needleman and Wunsch, 1970) and Levenstein distance (Levenstein, 1966) for more sensitive primer detection. Trimming of barcodes, linkers, and primers is done by a one step process after detecting all of them. By this process, PyroTrimmer can sort out sequence reads produced from the PCR amplicon mix of multiple genes by barcodes.

Trimming 3′ end primer sequence

Primer, linker and barcode sequences can be present in the 3′ end of reads if the sequence reads are long enough to cover the entire PCR amplicons. These sequences should be removed for accurate downstream analysis. The removal is however complicated by the fact that the primer sequences can also be found far upstream from the 3′ end of the sequencing reads. Therefore, PyroTrimmer examines the upstream region if the partial or full primer sequences exist using the same distance algorithms described above (Fig. 1). As sequencing quality decreases toward the end of reads, sequencing errors occur more frequently in the 3′ regions of reads. These errors can produce pairwise sequence alignments with multiple mismatches that decrease the sensitivity of primer detection. In order to reduce this problematic effect, PyroTrimmer compares the 3′ regions of reads with the last 10 bp of the primer sequences whose complementary bases in reads usually have quality scores higher than those of the upstream primer regions.

Trimming 3′ end with low quality score

Sequences usually have low quality at the 3′ end. Therefore, PyroTrimmer calculates the average quality value of nucleotides within the user-defined window size. The window starts at the 3′ end of a read, slides in the 5′ end direction until the average quality value is higher than the user-defined cutoff threshold and trims the downstream region of

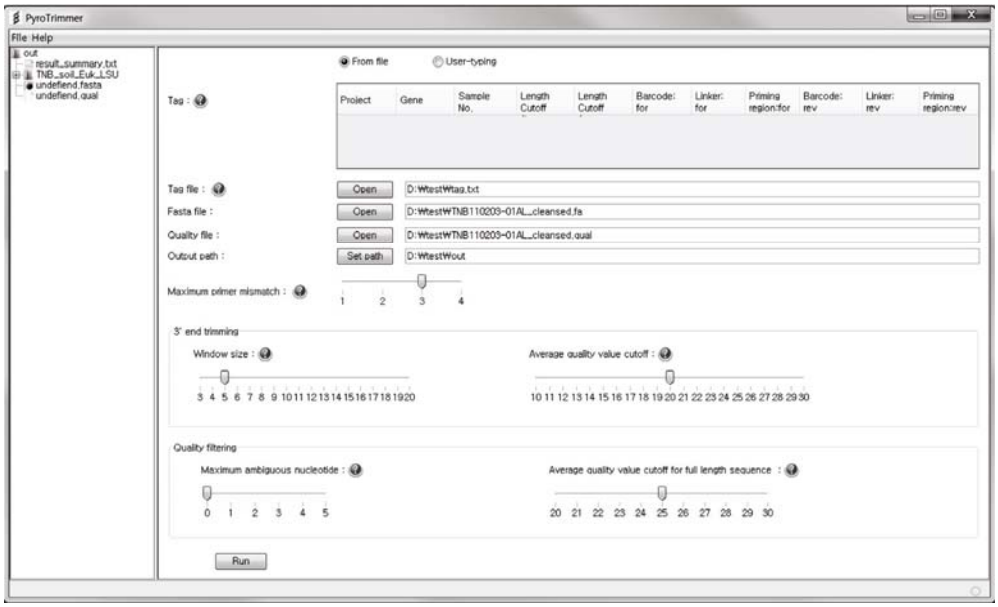

Fig. 2. PyroTrimmer graphic user interface.

the window (Fig. 1). The default cutoff value is 20 based on the Q20 test that is also used by the Genome Sequencer FLX Run Processor application to trim 3′-directed bases with low quality.

Filtering reads that have low average quality

Reads may also have low quality in the middle region. Such reads cannot be detected by the 3′ end quality trimming. In order to filter out these reads, PyroTrimmer calculates the average quality scores for individual reads and removes reads that have lower scores than the user-defined cutoff (Fig. 1).

Filtering sequence with ambiguous nucleotides

Sequence reads with ambiguous base N can also be removed. Although ambiguous base calls contribute marginally to the overall error rate, removal of sequence reads with the ambiguous bases can improve the global quality and reduce the error rate for the reads (Huse *et al.*, 2007; Gilles *et al.*, 2011).

Output

PyroTrimmer generates a number of output files. These include the trimmed result file, summary statistics file and files for discarded sequence reads (Fig. 1). Trimmed result files consist of FASTA sequences, quality scores and read length counts. When dealing with multiplex pyrosequencing data, this program produces the trimmed result files per gene of individual samples. Moreover, the trimmed reads that do not satisfy lower or upper bounds of the user-defined read lengths are saved in short or long read length files separately. The trimmed result files are saved in separate folders when they are results for different projects, which save the efforts of sorting out sequences for each research project. The summary statistics file contains a couple of statistics on the length distribution, read counts and others. Reads that are discarded are saved in the files of undefined FASTA sequences and quality scores.

Implementation

PyroTrimmer is implemented in JAVA and can be run on any kind of operating system without the requirement for additional programs to be installed. PyroTrimmer is available in both the command line and user-friendly GUI versions, which is easier to use. The GUI version consists of the following three panels (Fig. 2). The input panel accepts three types of input files: 1) FASTA sequences; 2) quality scores in FASTA format; and 3) tag information that includes sequences of barcodes, linkers and primers. There are two types of input widgets for tag information in the

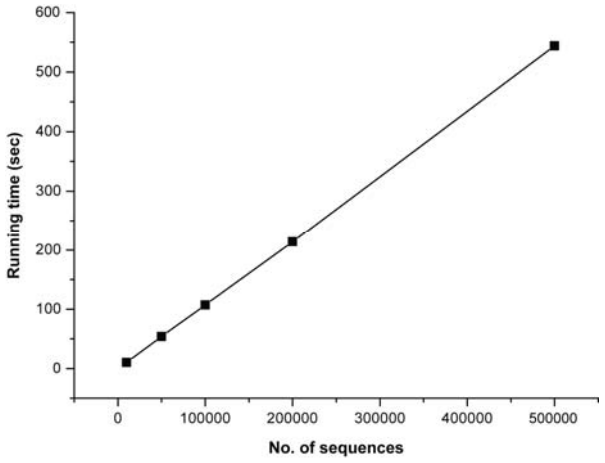

Fig. 3. Time complexity. The running time of PyroTrimmer was measured using the randomly sampled sequence datasets of size 10 K, 50 K, 100 K, 200 K, and 500 K. This program exhibits a linear computational complexity with the increases of dataset size.

Table 1. Comparison of PyroTrimmer with other applications

	Multi-sample	Multi-gene	Stand-alone GUI	Filtering read with base N	Trimming 3′ end primer	Filtering reads with low average quality	Trimming 3′ end with low quality
PyroTrimmer	Yes	Yes	Yes	Yes	Yes	Yes	Yes
RDP	Yes	No	No	Yes	Yes	Yes	No
TagCleaner	No	No	No	Yes	Yes	No	No
PyroTagger	Yes	Yes	No	Yes	No	Yes	No
CANGS	Yes	Yes	No	Yes	Yes	Yes	No

GUI version: Users can either copy and paste the tag information or import it from a file. This dual functionality allows easy handling of tag information. The output panel shows trimming results once the process is completed. Finally, the remaining panel has five input widgets for trimming and filtering out sequences. The GUI application also provides statistics on processed read counts and displays a progress bar in the bottom status panel for users to estimate the finish time.

Results and Discussion

PyroTrimmer has been tested on unpublished 16S ribosomal RNA sequence data of one million reads that were randomly extracted from the 454 reads of nearly 100 samples, each of which consists of around 10,000 reads. It took approximately 3 min for each sample on a Linux workstation with an Intel Xeon 2.66 GHz processor and 16 GB of RAM. We then randomly sampled sequences from the dataset of the one million reads and prepared the sequence datasets of 10 K, 50 K, 100 K, 200 K, and 500 K sizes. The datasets were tested on a Windows PC with an Intel i7 3.30 GHz processor and 8 GB of RAM. For all tests, PyroTrimmer showed the linear computational complexity with increases of data size (Fig. 3). Together with the fact that it does not utilize many computing resources while running, this shows that PyroTrimmer can efficiently operate on common machines. One can argue that the performance of PyroTrimmer needs to be compared with the existing programs. As described above, most of existing programs have multiple functionalities and are not dedicated to pre-processing sequence reads. Furthermore, the availability of trimming options varies depending on individual programs (Table 1). In this regard, the comparative exercise of PyroTrimmer with existing programs is not feasible under unified trimming conditions.

There are many pre-processing tools available for sequences produced from the 454 platforms. All the existing applications are either command line versions that are inconvenient to use and fairly complicated to install, or web-based versions that take a longer time to upload the high-throughput 454 data. In contrast, PyroTrimmer has the following advantages when compared with the other programs: 1) PyroTrimmer is available as a stand-alone GUI program and can be executed very easily and without the requirement for installation of other external programs. 2) PyroTrimmer trims and filters out low-quality reads accurately and efficiently. 3) PyroTrimmer supports processing of multiple samples and gives users more convenient data trimming options. Although PyroTrimmer has several advantages, the current version does not have a chimera filtering function, which is needed for pre-processing 454 reads. Future versions will be equipped with this added functionality.

Acknowledgements

We are sincerely grateful to Mr. Arshan Nasir for providing valuable comments. This research was supported by a grant from KRIBB Research Initiative Program, a grant from the Next-Generation BioGreen 21 Program, Rural Development Administration (PJ0090192012), and a grant from KOPRI Research Program (PE12030).

References

Cole, J.R., Wang, Q., Cardenas, E., Fish, J., Chai, B., Farris, R.J., Kulam-Syed-Mohideen, A.S., McGarrell, D.M., Marsh, T., Garrity, G.M., and *et al.* 2009. The Ribosomal Database Project: improved alignments and new tools for rRNA analysis. *Nucleic Acids Res.* **37**, D141–D145.

Gilles, A., Meglécz, E., Pech, N., Ferreira, S., Malausa, T., and Martin, J-F. 2011. Accuracy and quality assessment of 454 GS-FLX Titanium pyrosequencing. *BMC Genomics* **12**, 245.

Handelsman, J., Rondon, M.R., Brady, S.F., Clardy, J., and Goodman, R.M. 1998. Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. *Chem. Biol.* **5**, 245–249.

Huse, S.M., Huber, J.A., Morrison, H.G., Sogin, M.L., and Welch, D.M. 2007. Accuracy and quality of massively parallel DNA pyrosequencing. *Genome Biol.* **8**, R143.

Kunin, V. and Hugenholtz, B. 2010. PyroTagger. A fast, accurate pipeline for analysis of rRNA amplicon pyrosequence data. *The Open J.* **1**, 1.

Levenstein, V.I. 1966. Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics. Doklady.* **6**, 707–710.

Needleman, S.B. and Wunsch, C.D. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* **48**, 443–453.

Pace, N.R., Stahl, D.A., Olsen, G.J., and Lane, D.J. 1985. Analyzing natural microbial populations by rRNA sequences. *ASM News* **51**, 4–12.

Pandey, R.V., Nolte, V., and Schlötterer, C. 2010. CANGS: a user-friendly utility for processing and analyzing 454 GS-FLX data in biodiversity studies. *BMC Res. Notes* **3**, 3.

Schmieder, R., Lim, Y.W., Rohwer, F., and Edwards, R. 2010. TagCleaner: Identification and removal of tag sequences from genomic and metagenomic datasets. *BMC Bioinformatics* **11**, 341.

Thomas, T., Gilbert, J., and Meyer, F. 2012. Metagenomics – a guide from sampling to data analysis. *Microb. Inform. Exp.* **2**, 3.

Xu, J. 2006. Microbial ecology in the age of genomics and metagenomics: concepts, tools, and recent advances. *Mol. Ecol.* **15**, 1713–1731.