MOLECULAR ECOLOGY
RESOURCES

Molecular Ecology Resources (2014) **14**, 426–434 doi: 10.1111/1755-0998.12187

StreamingTrim 1.0: a Java software for dynamic trimming of 16S rRNA sequence data from metagenetic studies

G. BACCI,*† M. BAZZICALUPO,* A. BENEDETTI† and A. MENGONI*

**Department of Biology, University of Florence, via Madonna del Piano 6, Firenze I-50019, Italy, †Consiglio per la Ricerca e la Sperimentazione in Agricoltura, Centro di Ricerca per lo Studio delle Relazioni tra Pianta e Suolo (CRA-RPS), Via della Navicella 2/4, Roma I-00184, Italy*

Abstract

Next-generation sequencing technologies are extensively used in the field of molecular microbial ecology to describe taxonomic composition and to infer functionality of microbial communities. In particular, the so-called barcode or metagenetic applications that are based on PCR amplicon library sequencing are very popular at present. One of the problems, related to the utilization of the data of these libraries, is the analysis of reads quality and removal (trimming) of low-quality segments, while retaining sufficient information for subsequent analyses (e.g. taxonomic assignment). Here, we present StreamingTrim, a DNA reads trimming software, written in Java, with which researchers are able to analyse the quality of DNA sequences in fastq files and to search for low-quality zones in a very conservative way. This software has been developed with the aim to provide a tool capable of trimming amplicon library data, retaining as much as taxonomic information as possible. This software is equipped with a graphical user interface for a user-friendly usage. Moreover, from a computational point of view, StreamingTrim reads and analyses sequences one by one from an input fastq file, without keeping anything in memory, permitting to run the computation on a normal desktop PC or even a laptop. Trimmed sequences are saved in an output file, and a statistics summary is displayed that contains the mean and standard deviation of the length and quality of the whole sequence file. Compiled software, a manual and example data sets are available under the BSD-2-Clause License at the GitHub repository at https://github.com/GiBacci/StreamingTrim/.

Keywords: amplicon libraries, dynamic trimming, metagenetics, next-generation sequencing

Received 26 February 2013; accepted 7 October 2013

Introduction

The use of DNA barcoding coupled with recent improvements in next-generation sequencing is revolutionizing microbial ecology for the description of taxonomic composition and functionality of microbial communities. One of the most popular applications is the so-called metagenetic or barcode analysis (Creer *et al.* 2010; Huse *et al.* 2010; Sogin *et al.* 2006), which relies on sequencing PCR amplicons of a single gene of interest having taxonomic or functional value, for example the 16S rRNA gene for bacteria or the 18S rRNA for eucaryotes. Many studies have been performed using amplicon libraries to address issues in soil and aquatic ecology of prokaryotic and eucaryotic micro-organisms, as well as in meiofauna for recent examples, see (Bik *et al.* 2012; Creer & Sinniger 2012; Lecroq *et al.* 2011; Machida & Knowlton 2012a,b; Machida *et al.* 2012; Porazinska *et al.* 2010) and in host–microbe interactions (Andersson *et al.* 2008; Manter *et al.* 2010; Pini *et al.* 2012). The increasing interest in metagenetic approaches has been coupled with the development of new primers, strategies and computational tools to gain as much information as possible from generated amplicon libraries, trying to contain the costs of sequencing (Jones *et al.* 2011; Machida & Knowlton 2012a,b; Machida *et al.* 2012).

One of the most important problems related to the production and utilization of DNA sequence reads is the analysis of base quality and removal (trimming) of low-quality segments, while retaining sufficient information for subsequent analyses. Several trimming algorithms and software programs have been developed to cope with the clean-up of DNA sequence reads, for example SolexaQA DynamicTrim (solexaqa.sourceforge.net (Cox *et al.* 2010)), FASTX-ToolKit (http://hannonlab.cshl.edu/fastx_toolkit/, ConDeTri (http://code.google.com/p/condetri/ (Smeds & Kunstner 2011) and NGS QC Toolkit (Patel & Jain 2012). However, all these softwares have been developed to

Correspondence: Alessio Mengoni, Fax: +39-055-222565;
E-mail: alessio.mengoni@unifi.it

trim reads derived from genome sequencing projects, to prepare the reads for an assembly algorithm, or for mapping against a reference genome. On the contrary, in metagenetic analyses, it is very important to preserve the integrity of each read as much as possible, while reducing the loss of important taxonomic information present in a 16S or 18S rRNA gene sequence.

To overcome this limitation imposed by the existing trimming software programs, we have developed StreamingTrim using standard Java language and BioJava (Prlic *et al.*, 2012) libraries (included in the package). This software uses a very flexible 'dynamic window' algorithm (described below) to remove low-quality segments of DNA sequences, beginning from the end of each read in a sequences file. This approach is very useful because it allows users to set a more stringent quality cut-off, which increases the reads quality and reduces the risk of losing too much information. In addition, due to its graphical user interface, StreamingTrim can be simply installed and launched, allowing the software to be used even by inexperienced bioinformaticians, easily permitting 'wet lab' molecular ecologists to analyse their data. Here, we describe the program algorithm and the comparison with four other commonly used and published trimming software programs.

StreamingTrim workflow

StreamingTrim is a very conservative trimmer designed for improving the quality of fastq reads. This software is able to automatically decode ASCII-encoded quality, obtained with the most common sequencing techniques (Illumina, Solexa, Sanger and Solid). The quality of each base is checked against a quality cut-off parameter to identify and remove poor-quality bases. This cut-off parameter can either be set automatically by the software or manually specified by the user. If the user chooses to run StreamingTrim without any cut-off specification, the program performs a rapid statistical analysis and sets the cut-off automatically. The automatically derived cut-off value is calculated as the mean quality of all reads in the sequence file minus one quality standard deviation (e.g. if we have a file with a mean quality of 31.46 and a standard deviation of 6.54, the quality cut-off is set to 31.46–6.54 = 24.92 and approximated up to 25).

StreamingTrim algorithm workflows and example steps are reported in Fig. 1. Given a DNA sequence of length *N*, the algorithm starts from the last nucleotide (the *n*th nucleotide), using a window length (*W*) of 1, and checks if:

$$(\text{Quality}_{n^{\text{th}}} - \text{Cutoff}) \geq 0$$

If this is true, the algorithm will proceed by enlarging the window length by 1 (in this case putting *W* = 2), otherwise the *n*th nucleotide is removed. *N* is then decreased by the number of removed nucleotides (in this case 1), and *W* is set to 1. This process is repeated until the algorithm reaches the first nucleotide of the DNA sequence (*N* = 1), or if the trimmed sequence length goes below a minimum value previously chosen by the user (default 1). A formal description of the algorithm is shown here:

$$N = \text{sequence length};\ W = \text{window length};$$
$$M = (N - W)$$

$$T = \sum_{M<k\leq N} (\text{Nucl}_k - \text{Cutoff})$$

$$\text{If}\ T \geq 0 \rightarrow (W+1); \text{If}\ T<0 \rightarrow N = (N-W)\ \ W = 1$$

$$\text{Continue with the test T until}$$
$$N - W) \leq 0\ \text{or}\ N < \text{minimum length}$$

The above-reported algorithm has been developed to be as conservative as possible. In fact, a DNA segment is deleted only if all its nucleotides are considered to be of low quality. If there are only a few low-quality bases in a sequence, the segment is maintained to prevent loss of information. As stated earlier, the high parsimony of this software is very useful, especially in 'amplicon-based' analyses (barcode or metagenetic), for example those commonly employed in bacterial or fungal community analyses (16S or 18S rRNA analyses), in which even the loss of one nucleotide can be crucial to retain the taxonomic information.

Applications in the analysis of metagenetic libraries and comparison with existing software programs

Two different data sets, which contained 16S rRNA amplicons that were sequenced with two different techniques (Illumina and 454), were collected from the NCBI-SRA database to compare the performance of StreamingTrim with those of four other trimming algorithms (i.e. SolexaQA DynamicTrim (Cox *et al.* 2010), ConDeTri (Smeds & Kunstner 2011), NGS QC Toolkit (Patel & Jain 2012) and Mothur (Schloss *et al.* 2009)). The features of the two selected 'benchmarks' data sets are reported in Table 1.

First, we compared several software features that are not strictly algorithm-related (e.g. presence of a graphical user interface, reads type automatic recognition) (Table 2). Considering the computational time, 10 sub-data sets were generated with increasing number of sequences from both the 454 and the Illumina data sets. The obtained sub-data sets were used to generate a scatter plot of computation time over data set size (Fig. 2). The obtained results showed that StreamingTrim performed faster than

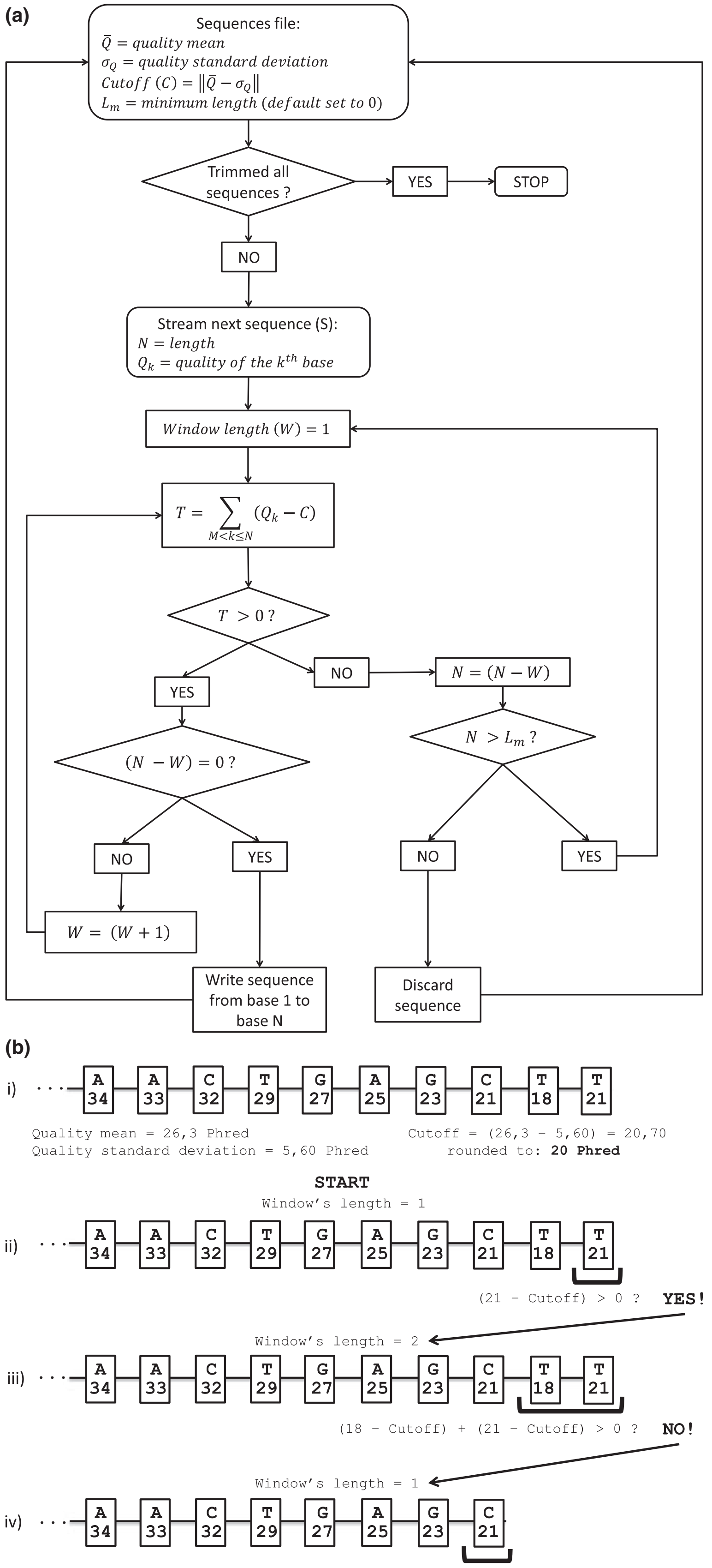
(a)
Sequences file:
$\bar{Q}$ = quality mean
σ_Q = quality standard deviation
Cutoff (C) = $\|\bar{Q} - \sigma_Q\|$
L_m = minimum length (default set to 0)
Trimmed all sequences ?
YES
STOP
NO
Stream next sequence (S):
N = length
Q_k = quality of the k^{th} base
Window length (W) = 1
$T = \sum_{M<k\leq N} (Q_k - C)$
T > 0 ?
YES
NO
N = (N − W)
(N − W) = 0 ?
$N > L_m$?
NO
YES
NO
YES
W = (W + 1)
Write sequence from base 1 to base N
Discard sequence
(b)
i)
A 34
A 33
C 32
T 29
G 27
A 25
G 23
C 21
T 18
T 21
Quality mean = 26,3 Phred
Quality standard deviation = 5,60 Phred
Cutoff = (26,3 - 5,60) = 20,70
rounded to: 20 Phred
START
Window's length = 1
ii)
(21 - Cutoff) > 0 ? YES!
Window's length = 2
iii)
(18 - Cutoff) + (21 - Cutoff) > 0 ? NO!
Window's length = 1
iv)

Fig. 1 Workflow of the StreamingTrim algorithm. (a) Flow chart of the StreamingTrim algorithm. (b) Schematic representation of StreamingTrim algorithm workflows. First (i), a sample sequence is selected from a sequence file with a mean quality of 26.30 Phred and a quality standard deviation (SD) of 5.60. Then (i), a quality cut-off is calculated by subtracting one SD from the quality mean. Next (ii), the last base of the sequence is analysed by subtracting the previously obtained cut-off from its quality value. If this result is bigger than 0, the base is maintained and (iii) the analysis window is increased by one. Now, the quality of each base is analysed as in step (ii), and the results are summed up. In the displayed example, the result is less than 0, and, consequently (iv), the two bases are removed from the sequence, and the size of the analysis window is set again to 1. All these steps are repeated until the sequence has been entirely analysed.

Table 1 Main features of the data sets used for testing trimmers*

Accession ID	Run/s ID/S	Sequencing technology	# Bases downloaded	Average reads length
SRX031467	SRR073271	Illumina	1.6 Gb	150
SRX060099	SRR040576 – SRR040991	454	1.1 Gb	570

*The ID of the accession, the sequencing technology and the number of bases with average read length is shown.

Table 2 Comparison of the main features of the trimmers used in the comparison

Features	Streaming Trim	SolexaQA DynamicTrim	ConDeTri	NGS QC Toolkit	Mothur
Trimming time for Illumina reads*	7 min/Gb	8 min/Gb	23 min/Gb	7 min/Gb	13 min/Gb
Trimming time for 454 reads*	11 min/Gb	6 min/Gb	25 min/Gb	4 min/Gb	11 min/Gb
Graphical user interface	Yes	No	No	No	Optional
Reads type recognition	Yes	Yes	No	Yes	No
Plots of results	Yes	Yes	No	No	No
Analysis of results	Yes	Yes	Yes	No	Yes

*Time values obtained from a regression analysis on the data shown in Fig. 2 (All *R* square values are bigger than 0.99, data not shown).

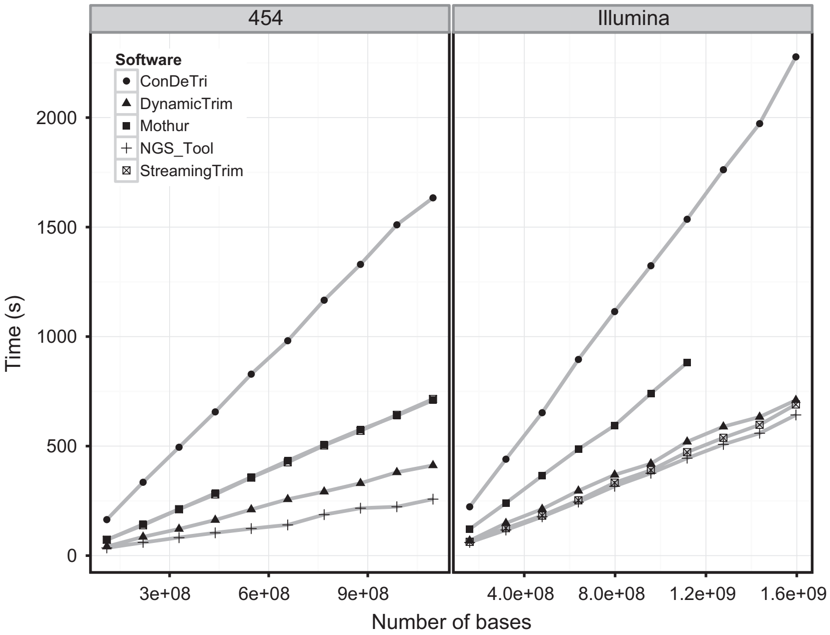

Fig. 2 Computational speed of the tested trimming software programs. Time plots were obtained by trimming 10 sequences files (sub-data sets), generated from Illumina and 454 data sets (20 samples total), with an increase in the number of sequences in each file. Each generated sub-data set file was processed, and the working time (expressed in seconds) of each trimming software program was recorded. The dimension of the sub-data sets was plotted against the working time of each software program. These analyses were performed on a standard desktop PC equipped with a 4-core CPU at 2.5 Ghz and 4 Gb RAM. Software names have been abbreviated as follows (abbreviated names are reported into parentheses): ConDeTri (ConDeTri), SolexaQA DynamicTrim (DynamicTrim), Mothur (Mothur), NGS QC Toolkit (NGS_Tool) and StreamingTrim (StreamingTrim).

CondeTri and very similar to Mothur with the 454 sub-data sets; with the Illumina sub-data sets, the performance of StreamingTrim was faster than CondeTri, Mothur and SolexaQA DynamicTrim and similar to NGS QC Toolkit. Finally, the whole data sets of Illumina and 454 reads were trimmed with all five trimming software programs using default settings (Figure S1).

To have a global view of the results, the mean quality and the number of bases of each data set were calculated using StreamingTrim before and after trimming with the five software programs; the obtained values are reported in Fig. 3. The data showed that StreamingTrim and NGS QC Toolkit are the most conservative trimmers (Fig. 3a), retaining up to 70% of the total bases. The mean quality of the data sets trimmed with StreamingTrim was increased, similar to the ones processed with the other four trimmers (Fig. 3b). However, if the user needs a more stringent analysis, StreamingTrim could be set to ensure even higher quality values, simply by increasing its quality cut-off parameter.

To compare the number of removed bases and the quality increment using a single metric, we introduced a trimming performance estimator, called *Z*-score. This estimator is proportional to the ratio between the

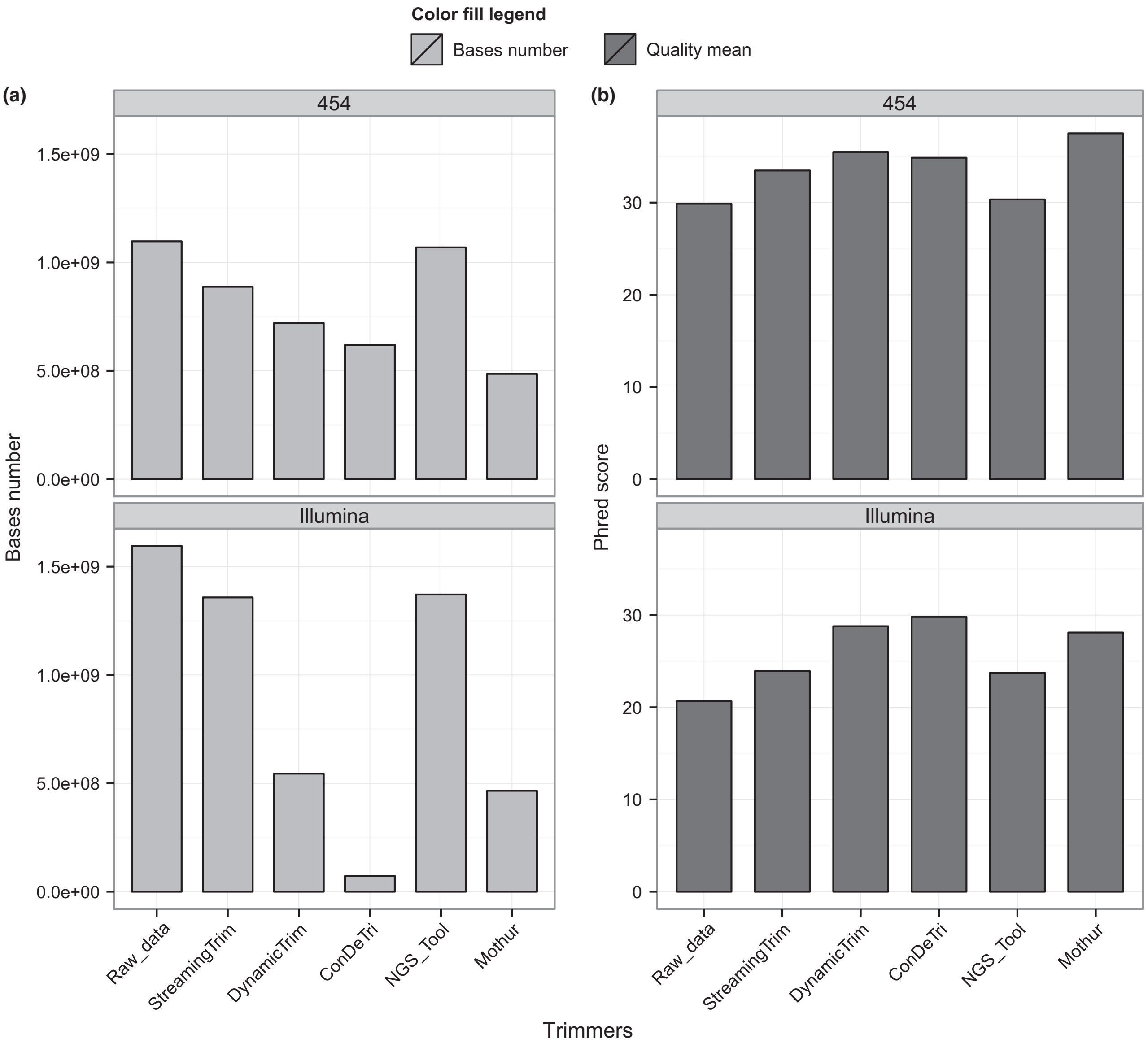

Fig. 3 Quality increase and base conservation of different trimming software programs. The number of bases (a) and the mean quality (b) in each data set before and after the trimming process are reported. 'Raw_data' values are those related to the pretrimming sequence file. Abbreviations for software names are as indicated in the legend of Figure 2.

increase in quality and the decrease in the number of bases for each data set. The *Z*-score was calculated as follows:

$$Z_{\text{score}} = \log_{10}\left(\frac{Q_{\text{diff}}}{|L_{\text{diff}}|}\right),$$

Where:

$$Q_{\text{diff}} = \frac{Q_f - Q_i}{Q_{\max} - Q_i} \text{ and } L_{\text{diff}} = \frac{L_f - L_i}{L_{\min} - L_i}$$

With:

Q_i = initial average quality; L_i = initial number of bases

Q_f = final average quality; L_f = final number of bases

$L_{\min}$ = minimum final number of bases (if users do not specify the minimum length parameter, this value is set to 0)

$Q_{\max}$ = maximaum final quality(for Phred Score this parameter is set to 40).

The results obtained with all tested trimming tools considered on the 454 and Illumina data sets showed that StreamingTrim had the highest *Z*-score values

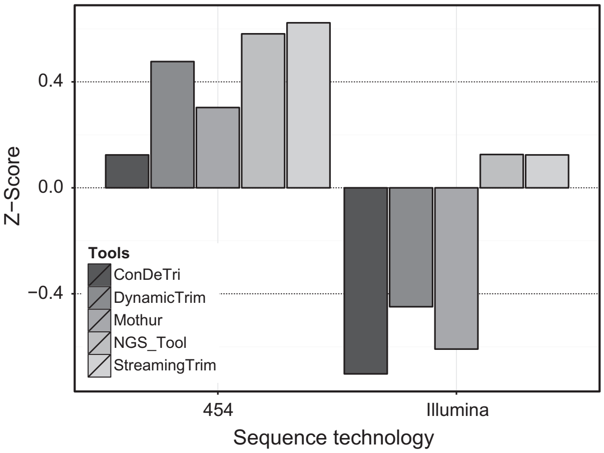

Fig. 4 *Z*-score of different trimming software programs. Bar charts of the *Z*-score after executing the trimming on two data sets (Illumina and 454) are shown. Negative values of the *Z*-score indicate that the percentage of bases lost during the trimming process is higher than the percentage of increase in quality. Positive values of the *Z*-score indicate that the quality increase is higher than the percentage of bases lost. Abbreviations for software names are as indicated in the legend of Figure 2.

(Fig. 4), indicating the presence of a good compromise between bases conservation and increase in reads quality. The only trimmer that showed a *Z*-score value similar to the one of StreamingTrim was NGS QC Toolkit.

Finally, the performance of all trimmers in terms of taxonomic profiling was compared. In fact, different trimming procedures may produce different taxonomic profiles due to different taxonomic assignments of trimmed 16S rRNA gene sequences (Huse *et al.* 2007). To compare different taxonomic profiles obtained with different trimmers, we have downloaded approximately 100 000 reads from a mock data set of an assemblage of different microbial species, present in the NCBI-SRA database (Haas *et al.* 2011; Jumpstart Consortium Human Microbiome Project Data Generation Working G 2012). In particular, we downloaded reads of a single experiment, obtained by sequencing a synthetic community containing equimolar concentrations of 16S rRNA genes for each species (even composition mock community – eMC). Details about the data set used are reported in Table 3.

First, the eMC data set was trimmed with the five trimming software programs, generating five different reads files. Next, each file was analysed using the RDP multiclassifier, a command line application derived from the RDP Classifier (Wang *et al.* 2007) that is able to analyse and compare multiple samples at the same time. Finally, a frequency table obtained from the RDP multiclassifier was processed using custom R scripts. The differences between the theoretical frequencies of genera present in the eMC and the actual frequencies obtained with the RDP multiclassifier are reported in Fig. 5. The results showed that all five trimming software programs produced very similar taxonomic patterns, indicating that StreamingTrim did not introduce taxonomic biases different from those of other trimming software programs. Regarding the fold change of taxonomic assignments compared with the theoretical ones, *Deinococcus* and *Staphylococcus* were over-represented, while *Bacillus*, *Enterococcus*, *Escherichia/Shigella*, *Helicobacter* and *Lactobacillus* were under-represented for all trimmers. The only genus that was drastically under-represented was the only archaeal member *Methanobrevibacter*, as already reported (Jumpstart Consortium Human Microbiome Project Data Generation Working, 2012). This discrepancy is due to the primers used, which target bacterial 16S rRNA sequences that diverge from the sequences found within Domain Achaea.

Although analysis of the trimmed files with the RDP multiclassifier resulted in similar taxonomic profiles, it produced different values in terms of (i) number of reads assigned to each genus present in eMC ('correctly assigned' reads); (ii) number of reads assigned to genera that are not present in the eMC ('wrongly assigned' reads) and (iii) number of unassigned reads at the genus level (Fig. 6). In particular, StreamingTrim provided the highest percentages of correctly assigned reads, together with Solexa QA DynamicTrim. As expected, stringent trimmers (such as ConDeTri and Mothur) removed approximately between 5 and 30% of the total reads trimmed, while less stringent trimmers (such as StreamingTrim, Solexa QA DynamicTrim and NGS QC Toolkit) retained up to 99% of trimmed reads.

Conclusions

StreamingTrim is a valuable trimming tool for analysing metagenetic data that can be easily used either by nonexpert bioinformaticians, due to its graphical user interface.

Table 3 Characteristics of the reads of the even composition mock community – eMC

Accession ID	Run/s ID/S	Sequencing technology	Number of bases downloaded	Average reads length
SRX021562*	SRR053856 – SRR053861	454	58 Mb	550

*Strains included in the eMC are reported in (Haas *et al.* 2011).

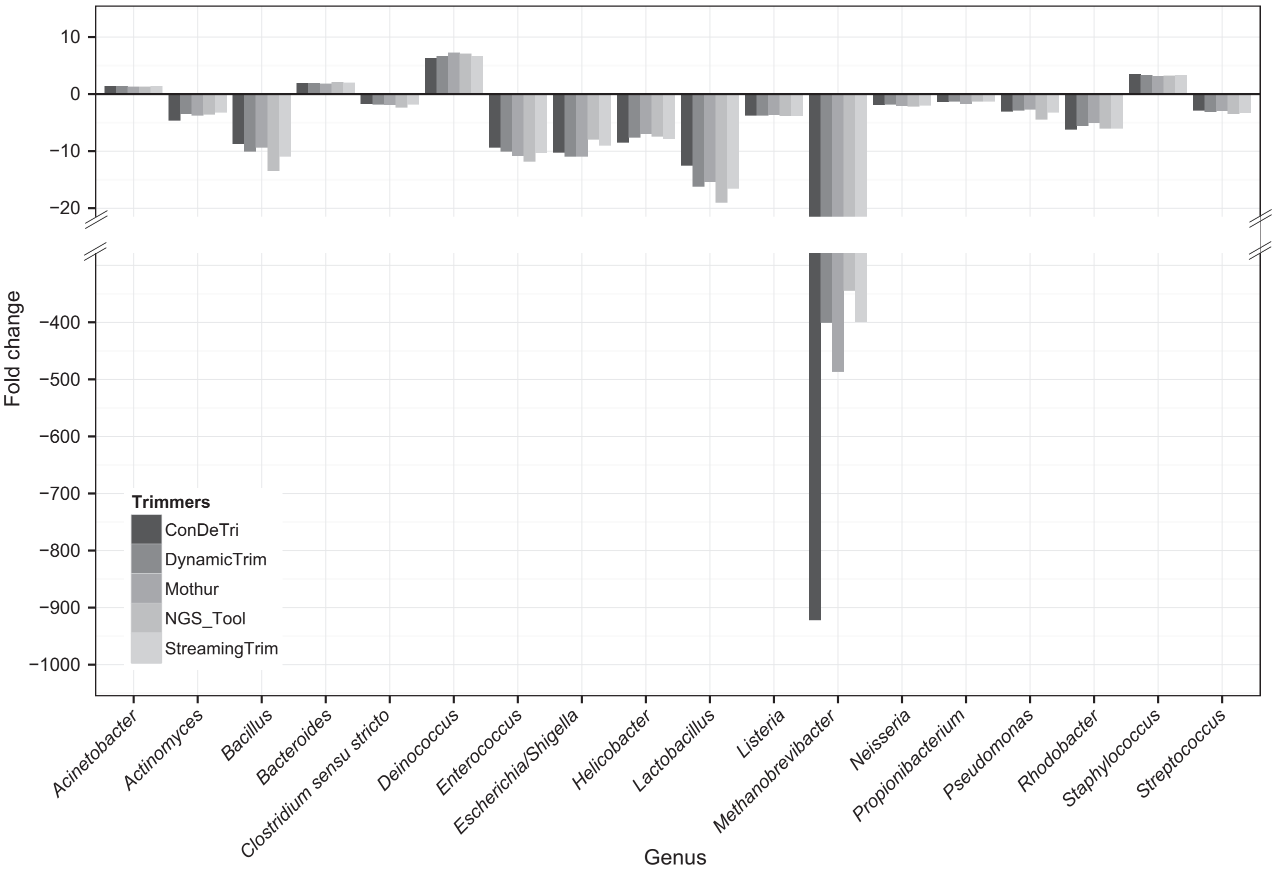

Fig. 5 Differences between theoretical and actual frequencies from the mock community (eMC) data set. Bar charts of differences between theoretical and observed frequencies of each genus in the eMC after trimming and assigning sequences with the RDP multi-classifier. Each difference obtained is expressed as fold change using relative frequencies of the synthetic community (Table S1). Formulas used are reported below:

${}^{eMC}_{i}$= relative frequency of genus i in the eMC

n_i^{obs} = relative frequency of genus i observed in the trimmed file

$$\frac{D_i^{rel} = n_i^{obs}}{n_i^{eMC}}.$$

Where D_i^{rel} is the ‘relative difference’ (fold change) between relative frequencies of genus i. Values of D_i^{rel} smaller than 1 have been replaced with the negative of their inverse (e.g. a value of 0.25 has been replaced with a value of −4). Software names abbreviations are defined in the legend of Fig. 2.

The main advantages of StreamingTrim over previously developed trimming software programs are based on flexible parameters settings, generation of trimmed sequences of high quality without loss of too much information, easy installation and portability on different computer platforms, due to the Java programming language. The length and quality of each sequence can be displayed in a plot before and after execution of the trimming process to provide an exhaustive overview of the results produced by the software. StreamingTrim can also convert fastq sequences to FASTA in order to analyse them without using other external software programs. Sequences can be converted after the trimming process or simultaneously to speed up the entire process.

The ratio between the increase in quality and the number of removed bases (*Z*-score) places StreamingTrim in

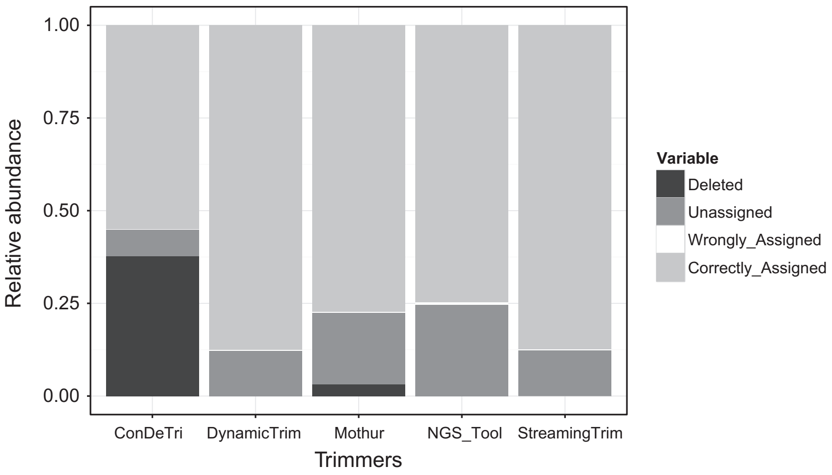

Fig. 6 Comparison of taxonomic assignments of reads of five trimming software programs. Bar chart of relative frequencies of removed sequences, unassigned sequences, correctly assigned sequences and wrongly assigned sequences for trimmed files obtained using all five software tools. Software name abbreviations are defined in the legend of Fig. 2.

a top position over previously developed trimmers. The decrease in the number of removed bases allows the user to perform detailed taxonomic analyses, while retaining as much information as possible. As a result, in the benchmark proposed here, StreamingTrim has correctly assigned a very high proportion of reads to their taxonomic classification, while performing better than several other software programs. In conclusion, StreamingTrim provides several improvements over previously developed trimming software programs and could be considered a valuable alternative or a complement, especially when used by 'wet lab' molecular ecologists.

Acknowledgements

This work was supported by a fellowship to GB from MIPAAF and by intramural funding of the university of Florence to AM and MB. We are also grateful to four anonymous reviewers who provided useful suggestions for improving the manuscript.

References

Andersson AF, Lindberg M, Jakobsson H, *et al.* (2008) Comparative analysis of human gut microbiota by barcoded pyrosequencing. *PLoS ONE*, **3**, e2836.

Bik HM, Sung W, De Ley P, *et al.* (2012) Metagenetic community analysis of microbial eukaryotes illuminates biogeographic patterns in deep-sea and shallow water sediments. *Molecular Ecology*, **21**, 1048–1059.

Cox M, Peterson D, Biggs P (2010) SolexaQA: at-a-glance quality assessment of Illumina second-generation sequencing data. *BMC Bioinformatics*, **11**, 485.

Creer S, Sinniger F (2012) Cosmopolitanism of microbial eukaryotes in the global deep seas. *Molecular Ecology*, **21**, 1033–1035.

Creer S, Fonseca VG, Porazinska DL, *et al.* (2010) Ultrasequencing of the meiofaunal biosphere: practice, pitfalls and promises. *Molecular Ecology*, **19**, 4–20.

Haas BJ, Gevers D, Earl AM, *et al.* (2011) Chimeric 16S rRNA sequence formation and detection in Sanger and 454-pyrosequenced PCR amplicons. *Genome Research*, **21**, 494–504.

Huse S, Huber J, Morrison H, Sogin M, Welch D (2007) Accuracy and quality of massively parallel DNA pyrosequencing. *Genome Biology*, **8**, R143.

Huse SM, Welch DM, Morrison HG, Sogin ML (2010) Ironing out the wrinkles in the rare biosphere through improved OTU clustering. *Environmental Microbiology*, **12**, 1889–1898.

Jones M, Ghoorah A, Blaxter M (2011) JMOTU and taxonerator: turning DNA barcode sequences into annotated operational taxonomic units. *PLoS ONE*, **6**, e19259.

Jumpstart Consortium Human Microbiome Project Data Generation Working G (2012) Evaluation of 16S rDNA-based community profiling for human microbiome research. *PLoS ONE*, **7**, e39315.

Lecroq B, Lejzerowicz F, Bachar D, *et al.* (2011) Ultra-deep sequencing of foraminiferal microbarcodes unveils hidden richness of early monothalamous lineages in deep-sea sediments. *Proceedings of the National Academy of Sciences of the United States of America*, **108**, 13177–13182.

Machida RJ, Knowlton N (2012a) PCR primers for metazoan nuclear 18S and 28S ribosomal DNA sequences. *PLoS ONE*, **7**, e46180.

Machida RJ, Knowlton N (2012b) Ways to mix multiple PCR amplicons into single 454 run for DNA barcoding. *Methods in molecular biology (Clifton, N.J.)*, **858**, 355–361.

Machida RJ, Kweskin M, Knowlton N (2012) PCR primers for metazoan mitochondrial 12S ribosomal DNA sequences. *PLoS ONE*, **7**, e35887.

Manter DK, Delgado JA, Holm DG, Stong RA (2010) Pyrosequencing reveals a highly diverse and cultivar-specific bacterial endophyte community in potato roots. *Microbial Ecology*, **60**, 157–166.

Patel RK, Jain M (2012) NGS QC Toolkit: a toolkit for quality control of next generation sequencing data. *PLoS ONE*, **7**, e30619.

Pini F, Frascella A, Santopolo L, *et al.* (2012) Exploring the plant-associated bacterial communities in *Medicago sativa* L. *BMC Microbiology*, **12**, 78.

Porazinska DL, Sung W, Giblin-Davis RM, Thomas WK (2010) Reproducibility of read numbers in high-throughput sequencing analysis of nematode community composition and structure. *Molecular Ecology Resources*, **10**, 666–676.

Prlic A, Yates A, Bliven SE, *et al.* (2012) BioJava: an open-source framework for bioinformatics in 2012. *Bioinformatics*, **28**, 2693–2695.

Schloss PD, Westcott SL, Ryabin T, *et al.* (2009) Introducing mothur: open source, platform-independent, community-supported software for describing and comparing microbial communities. *Applied and Environment Microbiology*, **75**, 7537–7541.

Smeds L, Kunstner A (2011) ConDeTri – a content dependent read trimmer for illumina data. *PLoS ONE*, **6**, e26314.

Sogin ML, Morrison HG, Huber JA, *et al.* (2006) Microbial diversity in the deep sea and the underexplored "rare biosphere". *Proceedings of the National Academy of Sciences*, **103**, 12115–12120.

Wang Q, Garrity GM, Tiedje JM, Cole JR (2007) Naive bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Applied and Environmental Microbiology*, **73**, 5261–5267.

The Research Group of Microbial Genomics at the Department of Biology, University of Florence, is interested in the ecology of soil bacterial communities and in the ecology of plant–microbe interactions, investigated through the use of massive sequencing and DNA barcoding. This work is part of the PhD programme of G.B., under the supervision of A.B. and A.M. G.B., A.M. and M.B. conceived the idea, G.B. developed the software and performed the testing, A.B. supervised the work, and A.M. and G.B. wrote the manuscript. All authors gave their consent to the final draft of the manuscript.

Data Accessibility

The program, user manual and example data set are available under the BSD-2-Clause License at the GitHub repository at https://github.com/GiBacci/StreamingTrim/.

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Figure S1 Quality plot of data sets before and after trimming.

Table S1 Analysis results for the mock community (eMC).

Data S1 StreamingTrim reference manual (PDF).