SeqTrim07: a pipeline for preprocessing sequence reads

Juan Falgueras

Dep. Lenguajes y Ciencias de la Computación,
Universidad de Málaga,
29071 Málaga, Spain
E-mail: juanfc@uma.es

Antonio J. Lara

Centro de Supercomputación y Bioinformática,
Universidad de Málaga,
29071 Málaga, Spain
E-mail: ajlara@uma.es

Guillermo Pérez-Trabado

Dep. Arquitectura de Computadores,
Universidad de Málaga,
29071 Málaga, Spain
E-mail: guille@ac.uma.es

Noé Fernández-Pozo,
Francisco R. Cantón and M. Gonzalo Claros*

Dep. Biología Molecular y Bioq.,
Universidad de Málaga,
29071 Málaga, Spain
E-mail: noefp@uma.es
E-mail: frcanton@uma.es
E-mail: claros@uma.es
*Corresponding author

Abstract: SeqTrim is a pipeline designed to preprocessing sequence reads. It is easy to install and configure, flexible even if default parameters are accurate for most purposes and usable as a web interface or a standalone command line application. It identifies the sequence insert by removing low quality sequences, cloning vector, poly-A or poly-T tails, adaptors and any contaminant sequence or unwanted feature. Several input and output formats are available, which enables its inclusion in already or newly defined sequence processing work flows. It outperforms preprocessors implemented in other web servers and standalone applications at least in detecting adaptors and chimeric clones. SeqTrim is under continuous refinement to deal with most sequence events due to collaboration between biologists and computer scientists.

Keywords: preprocessing; sequence reads; chromatograms; assembly; poly-A^+; poly-T^+; quality; web interface; command line; workflow; bioinformatics.

Reference to this paper should be made as follows: Falgueras, J., Lara, A.J., Pérez-Trabado, G., Fernández-Pozo, N., Cantón, F.R. and Claros, M.G. (2010) 'SeqTrim07: a pipeline for preprocessing sequence reads', *Int. J. Computational Intelligence in Bioinformatics and Systems Biology*, Vol. 1, No. 4, pp.370–382.

Biographical notes: J Falgueras is a Professor of Computer Science and Languages at Málaga University. His research interests include issues related to user interfaces and formal design. He has published a great number of articles at international journals, conference proceedings, book chapters, as well as open source software.

Antonio J. Lara is a PhD student developing new tools for sequence analysis. He has published a great number of articles at international journals, conference proceedings, book chapters, as well as open source software.

Guillermo Pérez-Trabado is a Professor of Computer Architecture at Málaga University. His research interests are related to parallelisation of processes and integration of databases with algorithms for bioinformatics. He has published a great number of articles at international journals, conference proceedings, book chapters, as well as open source software.

Noé Fernández-Pozo is a PhD student testing new tools for sequence analysis.

Francisco R. Cantón is a Professor of Biochemistry and Molecular Biology at Málaga University. His research interests include issues of expression analysis of genes related to wood synthesis in conifers. He has published a great number of articles at international journals, conference proceedings, book chapters, as well as open source software.

M. Gonzalo Claros is a Professor of Molecular Biology at Málaga University. His research interests are related to automated high-throughput analysis of gene expression data. He has published a great number of articles at international journals, conference proceedings, book chapters, as well as open source software.

1 Introduction

Sequencing projects and expressed sequence tags (ESTs) data have proven to be an important resource for gene discovery and mapping, and can be considered a backbone where to annotate eukaryotic and prokaryotic genomes by providing sequence information which identifies novel genes, gene location, polymorphisms and even intron-exon boundaries. The availability of automated sequencing has enabled the exponential growth rate of sequence data, although not always with the desired quality, mainly because EST are single reads of partially sequenced cDNA fragments and genomic DNA can contain ends of low quality sequence. This exponential growth requires efforts to be made in order to increase the quality and reliability of sequences

incorporated into databases, since up to 0.4% of sequences in nucleotide databases contain a high percent of nucleotides corresponding to contaminant sequences (Coker and Davies, 2004; Seluja et al., 1999). The situation is even worse in the EST databases, where vector contamination reaches 1.63% of sequences (Chen et al., 2007). Hence, improved bioinformatics tools are required to produce reliable preprocessing methods.

Preprocessing includes filtering of low-quality sequences, identification of specific features (like poly-A or poly-T tails, terminal transferase tails, adaptors, etc.), removal of contaminant sequences (from vector to any other artefacts) and trimming the undesired segments taking into account that all of these can be identified in single or multiple occurrences. There are some bioinformatics tools that serve to accomplish individual parts of the preprocessing (e.g., TrimSeq, TrimEST, VectorStrip, VecScreen, ESTPrep, crossmatch, Figaro), and other programs that deal with the complete preprocessing like PreGap4 (Bonfield et al., 1995) or the broadly used ones Lucy (Chou and Holmes, 2001; Li and Chou, 2004) or SeqClean (2008). Most of them need installation, are hard to configure, environment specific, or focused on specific needs (like a design only for ESTs), and it is not always possible to connect them with further processing tools for annotation or assembling. Moreover, use of such programs typically requires a change in implementation and design of either the program or the protocols within the laboratory itself. There are web implementations, ESTAnnotator (Hotz-Wagenblatt et al., 2003) or EST-pass (Lee et al., 2007) that start with preprocessing and go to assembling and/or annotating ESTs, but no web page is devoted exclusively to preprocessing.

This paper describes SeqTrim07, a software containing a flexible pipeline that deals with all preprocessing requirements (see above) for sequence reads. Its performance is compared with other broadly used open source applications devoted exclusively to preprocessing, like ESTPrep, Lucy2 and SeqClean.

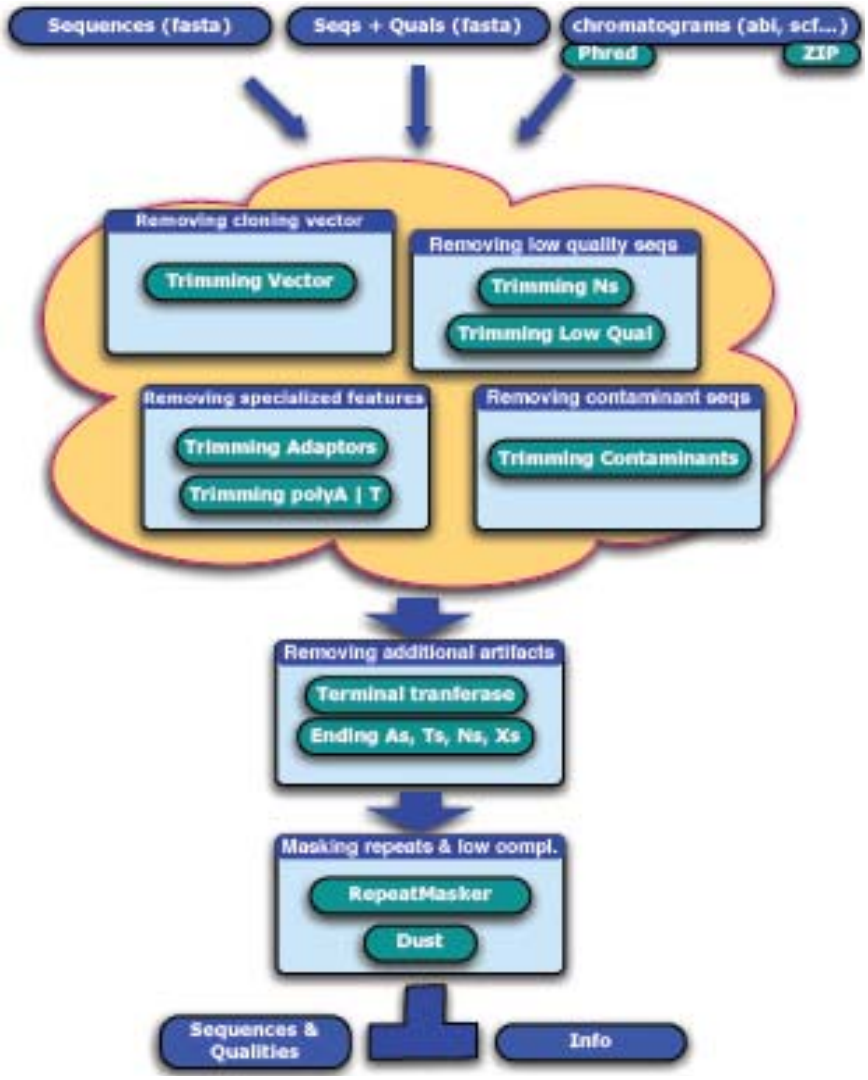
2 Implementation

SeqTrim07 has been programmed in Perl 5.8 using BioPerl (2008) libraries, and can be executed as a command line tool (ask for code to `jfalgueras@uma.es`) or as a web tool (`http://www.scbi.uma.es/seqtrim`). Command line version is more adequate for automatic batch processing or workflows while the web interface is more appropriate for user interactivity. SeqTrim makes use of the external programs phred (Ewing and Green, 1998; Ewing et al., 1998) for obtaining sequence and quality values, `BLAST` (Altschul et al., 1990) to compare sequences, and WU-`BLAST` to run RepeatMasker (2008) to mask repetitions and low complexity regions. All of them must be installed in the same machine than SeqTrim07, but are not included in the SeqTrim07 installer. It will work in any UNIX/Linux release, including OSX.

Uncompressing SeqTrim in `/usr/local` (or in any other directory defined in the `$PATH`) is enough to make it work. Configuration parameters in the `seqtrim` directory can be customisable transiently or permanently by the user. The file '`seqtrim.conf`' contains all configurable default parameters for analysis; parameters can be permanently modified according to the user needs editing the '`seqtrim.conf`', or changed for a single run via command-line options or the web interface. The seqtrim directory also contains the necessary databases, an editable file called '`RE_sites.txt`' that contains the usable restriction sites, and another editable file named '`adaptorSeqs.txt`' which contains a list of default adaptor sequences. Database modification is achieved

simply adding or removing sequences in `FASTA` format in the `seqtrim/DB` directory. Before each execution, SeqTrim verifies if there are changes in the databases and notifies it to the user, reformatting the database index when needed.

Figure 1 Detailed data-flow diagram of SeqTrim pipeline



Notes: It consists of four major steps (vector cleaning, specialised features, quality trimming, contamination removal) that can be executed in any order or skipped, and two ending steps (artefacts removal, and low complexity and repeats masking). The output is stored in a private area based on user's e-mail, and can be looked up asynchronously (see Figure 2).

The pipeline underlying SeqTrim07 goes throughout four independent and interchangeable processes plus two optional ending steps (Figure 1). A default pipeline order is provided, but users can change it completely, even skipping steps. After any process

but masking, only the excised sequence is passed to the next pipeline step. Pipeline description is given below following the default order.

3 Algorithm

3.1 Input and output

SeqTrim07 accepts several formats:

1 one or more FASTA file(s), each one containing one or more sequences

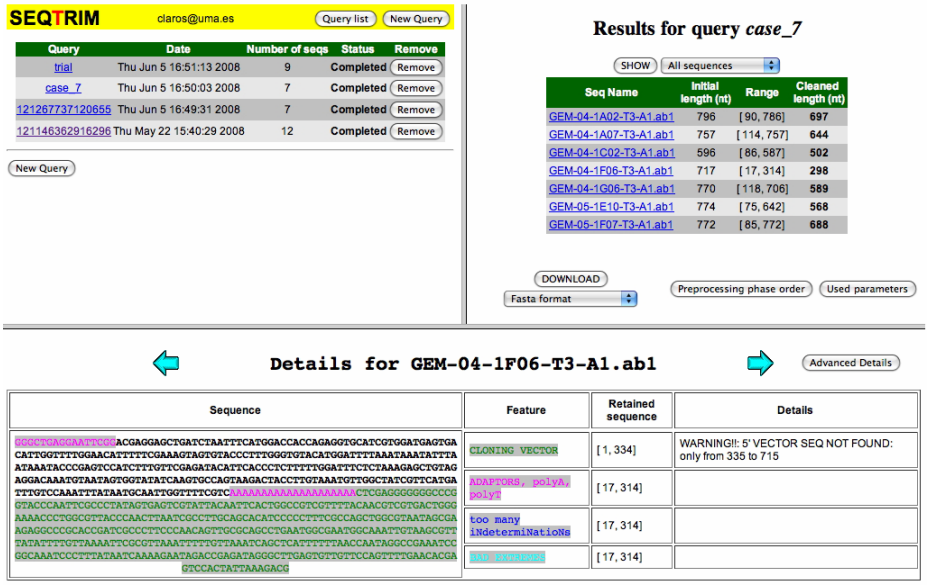
2 a FASTA file and the corresponding quality value file in FASTA format

3 a phd format file from phred

4 one chromatogram

5 a zip file containing a set of chromatograms.

Figure 2 Example of sequence trimming using different sets of sequences



Notes: The upper-left frame displays the status of all executions of SeqTrim made by the user ordered by date and time of run. The upper-right frame displays the list of the sequences analysed in a single query with original length, and range and length of the trimmed sequence. Buttons for saving result files and to know parameters and step order are also displayed. The bottom frame shows the original sequence coloured according to the different parts that have been removed. Legend on the right explains colour codes and a button to ask for more details are presented. Sequences to be detailed in the bottom frame are the same than listed in the upper-right frame.

Text files are directly processed, but when input sequences are chromatograms, the external program phred is called to obtain the quality value file. It must be understood that phred is not used at all to trim the low-quality end sequences. The first word in the description line of every input sequence is considered its name. Checks for sequence name duplications, as well as consistence between the sequence file and the quality value file (if provided), are performed.

Default output is a FASTA file containing only trimmed inserts. Alternative outputs, such as a text file containing user readable information concerning the trimming events for each sequence, a text file containing the names of the rejected sequences, or a FASTA file with masked sequences, can also be asked. Nucleotides whose quality value is not greater than the QV parameter (20 by default) are changed to lowercase. A coloured output of each sequence can also be seen on the screen, both using the command line or the web interface (Figure 2), which is intended to help users in the evaluation of preprocessing results. Results are stored in web server using user's e-mail as identifier.

3.2 Vector cleaning

The recommended first pipeline step relates to cloning vector detection comparing the NCBI's UniVec and the EMBL's emvec vector/adaptor libraries against each sequence using `BLAST` with relaxed parameters (`q -5`, `G 3`, `E 3`, `F 'm D'`, `e 1e-10`) to account for higher error rates at the beginning and end of reads so that users do not need to specify the cloning vector. `BLAST` alignment is parsed to identify regions that correspond to vector sequences, even if these regions are spliced into smaller DNA fragments that match in opposite orientation. SeqTrim07 is designed to locate cloning restriction sites only when cloning vector was not identified.

3.3 Removing specialised features

This step locates special features appearing in many sequences since they:

1 provide false sequences

2 mislead assembling or clustering algorithms that can be further used with these sequences

3 mislead researchers that use these contaminated sequences.

Adaptors, poly-A tails (only for ESTs) and poly-T tails (only for ESTs, which indicates that the sequence is in the reverse orientation) are considered unwanted features. Poly-A or poly-T detection is skipped if the input sequence corresponds to genomic DNA in order to gain CPU time.

Adaptors are located with blast2seq customising its parameters for short sequences (`W 7`, `F 'F'`, program '`blastn`'). Poly-A and poly-T tails are detected by specific heuristics developed by the authors, which includes the removal of one or more A at the 3' end of a sequence. In the case of ESTs, chimeric inserts are determined by the presence of two of these tails in the same sequence.

3.4 Quality trimming

Base-calling quality assessment for each nucleotide is taken into account to trim the original sequence in order to obtain the largest one with the highest quality. In cases where the input sequences are in a text FASTA format, where low quality nucleotides could be expressed by Ns (indetermination), SeqTrim07 can extract the largest subsequence without too many Ns. Parameters for determining the sequence quality can be changed. Since not all sequences include Ns, both trimming processes are split in order to enable users to skip the useless function. Trimming by quality values is only executed when input sequences are chromatograms or a quality value file is entered.

3.5 Contamination removal

During the experimental process, cloned sequences can result from contamination sources, such as DNA from *Escherichia coli* genome, cloning vector, cell plasmids, organelle, viruses, yeast, human, etc. Contamination is identified launching BLAST with trimmed sequences against a database of likely contaminants using default parameters and an expected cut-off fixed to 1e-3. SeqTrim07 is distributed with the genomes of *E. coli, S. cerevisiae*, lambda phage and several mitochondria. More databases, organised in a taxonomical way, are projected to be added in a near future. Length, position and identity of the contaminant sequence is provided for user information. Vector database is re-screened again at this moment, as well as adaptors, which serves to identify putative chimeras.

3.6 Removing other artefacts

This optional step is intended to be performed at the end of the preprocessing, since it is focused on removing any experimental artefact introduced in the apparently cleaned insert owed to molecular modifications. Extensions introduced by the terminal transferase enzyme, the Ns and/or Xs at both ends, and the Ts from the 5' end and/or As from the 3' end are screened by now.

3.7 Masking low complexity regions and repeats

This is the last step of the SeqTrim07 pipeline, in which the unwanted sequence is not removed but masked since low complexity regions and repeats are part of the real sequence, even if they can mistake further computer analysis. Low complexity regions due to nucleotide simple repeats are masked by an in-house algorithm. Repeats in nucleotide sequences are masked calling RepeatMasker using species-specific repeat libraries obtained from the Jurka (2000). The searching algorithm for RepeatMasker has been fixed to WU-BLAST in order to reduce the time spent in one sequence analysis.

3.8 Rejection criteria

A preprocessed sequence is rejected if it complies with one the following criteria:

1 there is an absence of insert between identified cloning vector boundaries

2 the usable sequence is not long enough (less than 100 bp by default)

3 there possibly are two inserts by the presence of two poly-A or poly-T tails, the concomitant presence of poly-A and poly-T tails, or an adaptor localisation in the middle of the sequence

4 the whole insert was masked.

We have seen that rejection when no cloning vector was found is not a good idea since sometimes there are rearrangements at the ends of cloning vectors that make them unrecognisable (Figure 2, bottom).

4 Results

Development of SeqTrim07 started in 1999, and along these years the algorithm has been trained with real data obtained from previous, different researches like EST from xylem tissue of *Pinus pinaster* (Cantón et al., 2003), EST from photosynthetic tissues of *Pinus sylvestris* (C Avila and FM Cánovas, unpublished results), SSH gene libraries from pine (DP Villalobos, S Díaz-Moreno, MG Claros and FR Cantón), or assembling BAC sequences (R Bautista, FR Cantón and MG Claros, manuscript in preparation).

Execution time for a single sequence will depend on the complexity of the given sequence. A total of 394 sequences were chosen for containing several preprocessable characteristics, with an average length of 755 nt. Running of SeqTrim07 (without masking) with such a training set on a 2.2 GHz Intel Core 2 duo processor provided the following results: 0.304 s for a single sequence and 29.19 s for a set of 96 sequences from a complete sequencing microplate. In a 1.6 GHz Itanium 2 processor it spends 0.48 s and 45.95 s, respectively. In following sections, comparison of SeqTrim07 performance with respect to other preprocessing software was assessed.

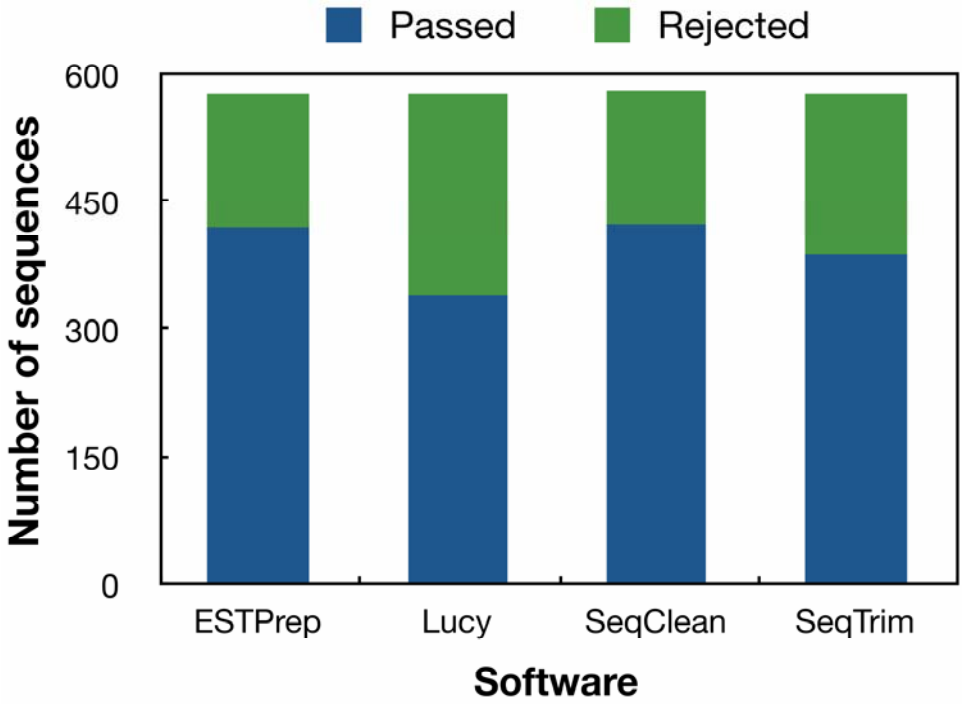
4.1 General preprocessing behaviour

SeqClean (2008) was chosen since it also uses `BLAST` to remove vectors, adaptors or primers and includes a call to RepeatMasker, which makes it very similar to SeqTrim07. Lucy (Chou and Holmes, 2001; Li and Chou, 2004) is another tested software which uses several base caller algorithms and an additional specific algorithm to preprocess sequences. Finally, the EST specific software ESTPrep (Scheetz et al., 2003) is also considered because it uses a heuristic match function to detect sequence features, and phred to obtain quality values. Although, cross-match is a restricted Smith-Waterman algorithm and has been incorporated into some EST processing packages, it has been discarded because it does not remove but masks vector-like regions, takes too much time to execute, and is not better than SeqClean or Lucy (Chen et al., 2007).

Not all testing programs are able to preprocess sequences different to ESTs. A collection of 576 EST chromatograms from the GEMINI project obtained in our laboratory (Cantón et al., 2003) were then used as the testing group. These reads resulted in 438.550 nucleotides, of which 53.8% were considered insert by ESTPrep, 37.4% by Lucy, 53.6% by SeqClean and 41.8% by SeqTrim07. The sequence reads had an average length of 761 nucleotides but, once preprocessed, the average insert size was 569 for ESTPrep, 490 for Lucy, 562 for SeqClean and 476 for SeqTrim07, clearly shown that SeqTrim renders the shorter sequences. Concerning to the number of passed/rejected

sequences (Figure 3), it seems that Lucy is the most restrictive, SeqTrim07 has an intermediate behaviour, and SeqClean and ESTPrep are the most permissive with similar results. Equivalent outcomes were derived with the number of sequences instead of the number of nucleotides.

Figure 3 Overview of the behaviour concerning rates of sequence acceptation or rejection by each tested algorithm



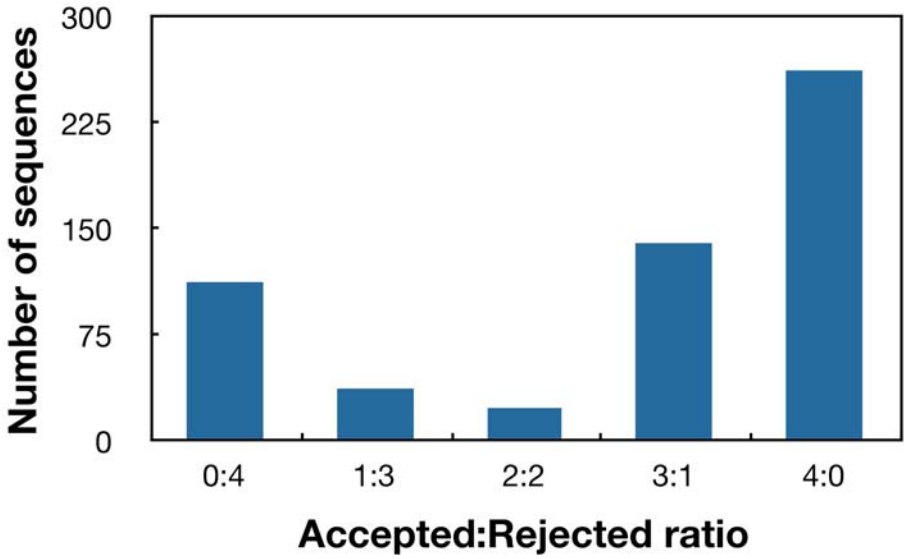
4.2 Concordance between algorithms

Concordance among the algorithms was tested by assessing the number of sequences accepted and rejected. The four softwares agree in 375 sequences [113 of 0:4 plus 262 of 4:0 in Figure 4(a)] which corresponds to 64.1% of sequences. If the agreement is relaxed to three algorithms, the concordance increases to 95.8%. SeqTrim07 is mainly consistent with ESTPrep and SecClean [92.0% and 93.9%, respectively, Figure 4(b)], which are also consistent among them (92.0%), but not with Lucy (68.7%) which clearly disagrees with ESTPrep and SecClean (70.5% and 72.4%, respectively). Since SeqClean and ESTPrep are routinely utilised by institutions devoted to sequence analysis, their agreement with SeqTrim07 supports its performance. It should also be noted that SeqTrim07 provides the shortest final sequence in 217 cases, the second to last in 42 cases, the third to last only in three cases, and it never provides the longest one.

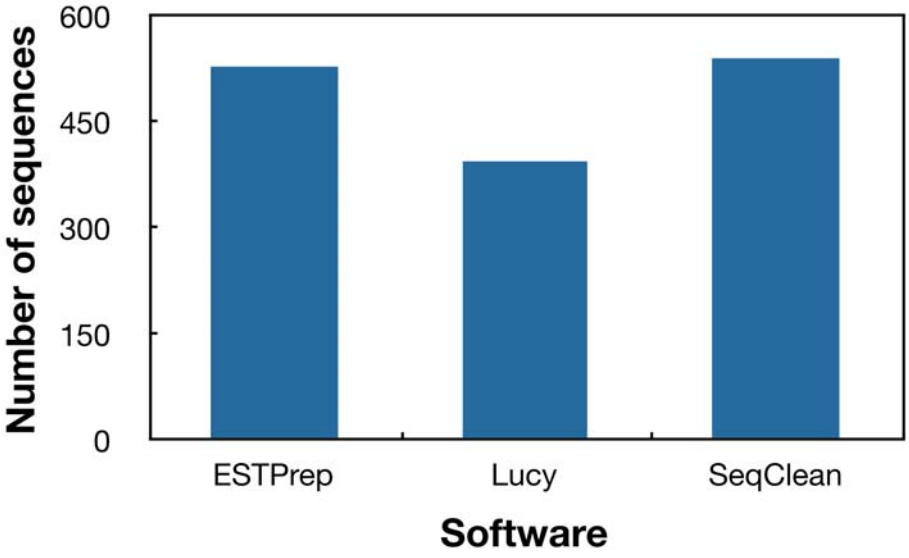
Finally, the agreement between the most coincident softwares (SeqTrim07 and SeqClean) was assessed in a crossed analysis. When sequences trimmed by SeqTrim07 were used as input for SeqClean, no changes were observed, indicating that both softwares remove the same features. On the contrary, when sequences trimmed by

SeqClean were used as input for SeqTrim07, most of them were slightly shorter mainly due to adaptor removal that SeqClean did not detect, although sometimes differences were related to low quality sequences that were not removed by SeqClean (results not shown). Putative chimeric sequences are also discarded with SeqTrim07 but not with SeqClean.

Figure 4 Concordance of SeqTrim07 with ESTPrep, Lucy and SeqClean, (a) degree of agreement from concordance in rejection (0:4, which means none of the four implementations reject a sequence) to concordance in acceptation (4:0, which means that the four accept the same sequence) (b) agreement of SeqTrim07 with the other three



(a)



(b)

5 Discussion

Even if there are many DNA preprocessing algorithms in the bioinformatics literature, getting them to work correctly may be very hard, and require an extra programming effort to consider unlikely special cases that appear when handling large amounts of sequences or the data quality are very low (Chou and Holmes, 2001). Hence, a new software was needed. Due to collaborations between computer scientists and biologists at the University of Málaga, the distance between the theoretical design of a bioinformatics solution and the successful implementation of the solution because of the inherently unpredictable nature of biological data, has been shorten. The product of such a collaboration, SeqTrim07, is under continuous development since its creation in 1999, which makes it well suited for the needs of large-scale sequence preprocessing, providing a time- and cost-effective solution, and for dealing with all potential events that can occur in a sequencing project. The use of SeqTrim07 has significantly reduced the time and complexity involved in a number of gene discovery projects at the University of Málaga. Unlike other equivalent software, installation of SeqTrim07 does not require special knowledges since there is nothing to compile and only requires installation of freely available `BLAST`, WU-`BLAST`, bioperl libraries, phred and RepeatMasker. Configuration files and databases provided with SeqTrim07 can be customised, although most parameters will never need a change. In fact, SeqTrim07 offers more customisation than SeqClean or ESTPrep, nearly pregap4, but do not present the more than the overwhelming 40 parameters that can be modified in Lucy (Li and Chou, 2004).

Sequence preprocessors are not expected to be used alone but in a pipeline with other programs (Liang et al., 2000; Masoudi-Nejad et al., 2006; Miller et al., 1999; Scheetz et al., 2003). Sometimes, constructing a pipeline is not easy mainly due to input/output formats or any other program peculiarities. This has been considered in SeqTrim07, since its flexibility regarding input and output formats contrasts with other sequence preprocessors that admit only one single type of sequence file: SeqClean, ESTPass or ESTPrep accept `FASTA` sequences while Lucy and pre-gap4 accept any chromatogram. Concerning the output, saving final sequences as trimmed or masked sequences enables the possibility to easily include SeqTrim07 in other previous workflows like pred/crossmatch/repeatmasker/phrap, or EST2UNI (Forment et al., 2008).

Most sequence preprocessors must be used only as command line programs (SeqClean, ESTPrep, TrimSeq, phred/crossmatch, Figaro), only as web pages (VecScreen, EMBVec Query) or as command line and a GUI interface (Lucy, pregap4). However, SeqTrim07 is designed for any kind of skilled user as a web or standalone application; no other website is devoted exclusively to preprocessing since preprocessing use to be included in more general pipelines [e.g., EGAssembler (Masoudi-Nejad et al., 2006) and ESTPass (Lee et al., 2007)]. SeqTrim07 uses coloured output for differentiation of trimmed regions in each sequence (Figure 2) to facilitate result interpretation as well as comparison between cleaned and original sequences, since they are on the same string instead of two synchronised scrolling panels as occurs in Lucy.

SeqTrim differential behaviour can be explained as follows:

1 SeqTrim07 can remove adaptor sequences that are not included in SeqClean or ESTPrep. Although these ones are able to finally remove adaptors if they are attached to the contamination database, this is not an easy task for unskilled users.

2 The parameters introduced to remove the low quality sequences are the most restrictive. This explains why SeqTrim07 provided the shortest sequences, and that more sequences are then rejected by shortness reasons.

3 Unlike others, SeqTrim07 is able to remove the chimeric inserts whether two poly-A and/or poly-T are found in the sequence or whether an adaptor is detected inside the trimmed sequence|it should be noted that the only pipeline that declares to remove double inserts is ESTPass (Lee et al., 2007), but uses a quite different approach that, in our hands, mark as chimeric EST sequences that were not.

4 SeqTrim07, tries to align vector, adaptor and restriction site sequences to localise insert instead of only one criterium. This seems to be better than locating its boundaries as a function of restriction site presence (Scheetz et al., 2003), since cloning sites can have experimental or base-calling errors and are too short to provide certainty, or locating it simply as a vector alignment (SeqClean, 2008).

5 Most preprocessing pipelines start by removing indeterminations and low quality regions, but we have found that vector and adaptor location is more reliable using the whole sequence and then applying the quality removal.

Moreover, starting with low quality removal may remove key information like sequence orientation or poly-A or poly-T presence. Recently, it has been published that base callings use to remove sequences that can be identified as vector (White et al., 2008).

Acknowledgements

The authors would like to acknowledge the Spanish MEC for supporting grants AGL2009-12139-C02-02 and BIO2009-07490 as well as the Junta de Andalucía for the grant AGR-663 and founding to the research groups CVI-114, TIC-160 and TIC-113.

References

Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. (1990) 'Basic local alignment search tool', *J. Mol. Biol.*, Vol. 215, pp.403–410.

BioPerl (2008) available at http://www.bioperl.org/ (accessed on 20 February).

Bonfield, J.K., Smith, K. and Staden, R. (1995) 'A new DNA sequence assembly program', *Nucleic Acids Res*., Vol. 23, pp.4992–4999.

Cantón, F., Le Provost, G., Garcia, V., Barré, A., Frigerio, J.M., Paiva, J., Fevereiro, P., Avila, C., Mouret, J.F., de Daruvar, A., Cánovas, F. and Plomion, C. (2003) 'Transcriptome analysis of wood formation in maritime pine', in Espinel, S., Barredo, Y. and Ritter, E. (Eds.): *Sustainable Forestry, Wood Products and Biotechnology*, DFA-AFA Press, Vitoria-Gasteiz.

Chen, Y-A., Lin, C-C., Wang, C-D., Wu, H-B. and Hwang, P-I. (2007) 'An optimized procedure greatly improves EST vector contamination removal', *BMC Genomics*, Vol. 8, p.416.

Chou, H.H. and Holmes, M.H. (2001) 'DNA sequence quality trimming and vector removal', *Bioinformatics*, Vol. 17, pp.1093–1104.

Coker, J.S. and Davies, E. (2004) 'Identifying adaptor contamination when mining DNA sequence data', *Biotechniques*, Vol. 37, pp.194–198.

Ewing, B. and Green, P. (1998) 'Base-calling of automated sequencer traces using phred. II. Error probabilities', *Genome Res*., Vol. 8, pp.186–194.

Ewing, B., Hillier, L.,Wendl, M.C. and Green, P. (1998) 'Base-calling of automated sequencer traces using phred. I. Accuracy assessment', *Genome Res*., Vol. 8, pp.175–185.

Forment, J., Gilaber, F., Robles, A., Conejero, V., Nuez, F. and Blanca, J.M. (2008) 'EST2uni: an open, parallel tool for automated EST analysis and database creation, with a data mining interface and microarray expression data integration', *BMC Bioinformatics*, Vol. 9, p.5.

Hotz-Wagenblatt, A., Hankeln, T., Ernst, P., Glatting, K.H., Schmidt, E.R. and Suhai, S. (2003) 'ESTAnnotator: a tool for high throughput EST annotation', *Nucleic Acids Res*., Vol. 31, pp.3716–3719.

Jurka, J. (2000) 'Repbase update: a database and an electronic journal of repetitive elements', *Trends Genet*, Vol. 16, pp.418–420.

Lee, B., Hong, T., Byun, S.J., Woo, T. and Choi, Y.J. (2007) 'ESTpass: a web-based server for processing and annotating expressed sequence tag (EST) sequences', *Nucleic Acids Res.*, Vol. 35, pp.W159–W162.

Li, S. and Chou, H.H. (2004) 'LUCY2: an interactive DNA sequence quality trimming and vector removal tool', *Bioinformatics*, Vol. 20, pp.2865–2866.

Liang, F., Holt, I., Pertea, G., Karamycheva, S., Salzberg, S. and Quackenbush, J. (2000) 'An optimized protocol for analysis of EST sequences', *Nucleic Acids Res*., Vol. 28, pp.3657–3665.

Masoudi-Nejad, A., Tonomura, K., Kawashima, S., Moriya, Y., Suzuki, M., Itoh, M., Kane-hisa, M., Endo, T. and Goto, S. (2006) 'EGassembler: online bioinformatics service for large-scale processing, clustering and assembling ESTs and genomic DNA fragments', *Nucleic Acids Res*., Vol. 34, pp.W459–W462.

Miller, R.T., Christoffels, A.G., Gopalakrishnan, C., Burke, J., Ptitsyn, A.A., Broveak, T.R. and Hide, W.A. (1999) 'A comprehensive approach to clustering of expressed human gene sequence: the sequence tag alignment and consensus knowledge base', *Gemone Res*., Vol. 9, pp.1143–1155.

RepeatMasker (2008) available at http://www.repeatmasker.org (accessed on 20 February).

Scheetz, T.E., Trivedi, N., Roberts, C.A., Kucaba, T., Berger, B., Robinson, N.L., Birkett, C.L., Gavin, A.J., O'Leary, B., Braun, T.A., Bonaldo, M.F., Robinson, J.P., Shefield, V.C., Soares, M.B. and Casavant, T.L. (2003) 'ESTprep: preprocessing cDNA sequence reads', *Bioinformatics*, Vol. 19, pp.1318–1324.

Seluja, G.A., Farmer, A., McLeod, M., Harger, C. and Schad, P.A. (1999) 'Establishing a method of vector contamination identification in database sequences', *Bioinformatics*, Vol. 15, pp.106–110.

SeqClean (2008) Available at http://compbio.dfci.harvard.edu/tgi/software/ (accessed on 20 February).

White, J.R., Roberts, M., Yorke, J.A. and Pop, M. (2008) 'Figaro: a novel statistical method for vector sequence removal', *Bioinformatics*, Vol. 24, pp.462–467.