BIOINFORMATICS APPLICATIONS NOTE

Vol. 28 no. 3 2012, pages 433–435
doi:10.1093/bioinformatics/btr669

Sequence analysis

Advance Access publication December 6, 2011

Identification and removal of ribosomal RNA sequences from metatranscriptomes

Robert Schmieder[1,2,*], Yan Wei Lim[3] and Robert Edwards[1,4,*]

[1]Department of Computer Science [2]Computational Science Research Center [3]Department of Biology, San Diego State University, San Diego, CA 92182 and [4]Mathematics and Computer Science Division, Argonne National Laboratory, Argonne, IL 60439, USA

Associate Editor: Ivo Hofacker

ABSTRACT

Summary: Here, we present riboPicker, a robust framework for the rapid, automated identification and removal of ribosomal RNA sequences from metatranscriptomic datasets. The results can be exported for subsequent analysis, and the databases used for the web-based version are updated on a regular basis. riboPicker categorizes rRNA-like sequences and provides graphical visualizations and tabular outputs of ribosomal coverage, alignment results and taxonomic classifications.

Availability and implementation: This open-source application was implemented in Perl and can be used as stand-alone version or accessed online through a user-friendly web interface. The source code, user help and additional information is available at http://ribopicker.sourceforge.net/.

Contact: rschmied@sciences.sdsu.edu; redwards@cs.sdsu.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on September 5, 2011; revised on November 28, 2011; accepted on November 29, 2011

1 INTRODUCTION

Metatranscriptomic approaches are drastically improving our understanding of metabolism and gene expression in microbial communities. By investigating all functional mRNA transcripts isolated from an environmental sample, metatranscriptomic analyses provide insights into the metabolic pathways important for a community at the time of sampling. Although metatranscriptomes are used to investigate metabolic activities, the majority of RNA recovered in metatranscriptomic studies is ribosomal RNA (rRNA), often exceeding 90% of the total reads (Stewart *et al*., 2010). Even after various treatments prior to sequencing, the observed rRNA content decreases only slightly (He *et al*., 2010) and metatranscriptomes still contain significant amounts of rRNA.

Although rRNA-like sequences are occasionally removed from metatranscriptomes, the removal is performed only with a subset of the publicly available rRNA sequences. Failure to remove all rRNA sequences can lead to misclassifications and erroneous conclusions during the downstream analysis. It is estimated that misannotations of rRNA as proteins may cause up to 90% false positive matches of rRNA-like sequences in metatranscriptomic studies (Tripp *et al*., 2011). The potential for false positives arrises from a failure to completely remove all rRNA prior to translating the putative rRNA and querying a protein database. The rRNA operons in Bacteria and Archaea are not known to contain expressed protein coding regions that at the same time code for rRNA and therefore, annotations of proteins in rRNA coding regions should be presumed to be misannotations (Aziz *et al*., 2008). Metagenomic sequence data generated to asses the metabolic potential of a community will also be affected by false positive matches of rRNA sequences when querying a protein database. Therefore, transcript analysis should only proceed after it has been verified that all rRNA-like sequences have been found and removed from the dataset to allow accurate identification of the transcribed functional content. The high-throughput nature of community sequencing efforts necessitates better tools for the automated preprocessing of sequence datasets.

Here, we describe an application able to provide graphical guidance and to perform identification, classification and removal of rRNA-like sequences on metatranscriptomic data. The application incorporates a modified version of the BWA-SW program (http://bioinformatics.oxfordjournals.org/citmgr?gca=bioinfo;26/5/589), and is publicly available through a user-friendly web interface and as stand-alone version. The web interface allows online analysis using rRNA sequences from public databases and provides data export for subsequent analysis.

2 METHODS

2.1 Implementation and computational platform

The riboPicker application was implemented as stand-alone and web-based version in Perl. The web application is currently running on a web server with Ubuntu Linux using an Apache HTTP server to support the web services. The alignments are computed on a connected computing cluster with 10 working nodes (each with 8 CPUs and 16 GB RAM) running the Oracle Grid Engine version 6.2. All graphics are generated using the Cairo graphics library (http://cairographics.org/).

2.2 Identification of rRNA-like sequences

The identification of rRNA-like sequences is based on sequence alignments using a modified version of the BWA-SW program. The modifications do not change the default behavior of the algorithm and include parameter forced changes in the alignment of ambiguous bases and the generation of an alternative output. The documentation provides a detailed list of changes and is available on the program website.

riboPicker uses query sequence coverage, alignment identity and minimum alignment length thresholds to determine if an input sequence

*To whom correspondence should be addressed.

is an rRNA-like sequence or not. This approach is based on the idea that looking for similar regions consists of grouping sequences that share some minimum sequence similarity over a specified minimum length. Threshold percentage values are rounded toward the lower integer and should not be set to 100% if errors are expected in the input sequences. The results for multiple databases are automatically joined before generating any outputs.

Using simulated datasets, we evaluated the classification of rRNA-like sequences and showed that riboPicker performed with high accuracy comparable to the latest version of meta_rna (Huang *et al*., 2009) and BLASTn (Supplementary Material). A comparison on real metatranscriptomic data showed that riboPicker processes data more than twice as fast as Hidden Markov Model (HMM)-based programs and >100 times faster than BLASTn (Supplementary Material).

2.3 Reference databases

The web-based version offers preprocessed databases for 5S/5.8S,16S/18S and 23S/28S rRNA sequences from a variety of resources, currently including SILVA (Pruesse *et al*., 2007), RDP (Cole *et al*., 2009), Greengenes (DeSantis *et al*., 2006), Rfam (Gardner *et al*., 2011), NCBI (Sayers *et al*., 2011) and HMP DACC (The NIH HMP Working Group *et al*., 2009). To reduce the number of possibly misannotated entries, sequences were filtered by length to remove very short and long sequences and by genomic location to remove overlapping rRNA misannotations. The remaining sequences were then converted into DNA sequences (if required) and filtered for read duplicates to reduce redundancy in the sequence data. Detailed information for each reference database is provided on the website. Taxonomic information was either retrieved with the sequence data from the resources or was added based on the NCBI Taxonomy. The databases are automatically updated on a regular basis and can be requested from the authors for offline analysis. A non-redundant database is made available for the stand-alone version on the program website.

3 WEB-INTERFACE

3.1 Inputs

The web interface allows the submission of compressed FASTA or FASTQ files to reduce the time of data upload. Uploaded data can be shared or accessed at a later point using unique data identifiers. It should be noted at this point that the input datasets should only contain quality-controlled, preprocessed sequences to ensure accurate results (Schmieder and Edwards, 2011). In addition to the sequence data, the rRNA reference databases have to be selected from the list of available databases.

Unlike the stand-alone version, the web-based program allows the user to define threshold parameters based on the results after the data are processed. This does not require an *a priori* knowledge of the best parameters for a given dataset and the parameter choice can be guided by the graphical visualizations.

3.2 Outputs

Users can download the results in FASTA or FASTQ (if provided as input) format or its compressed version. Results will be stored for the time selected by the user (either 1 day or 1 week), if not otherwise requested, on the web server using a unique identifier displayed during data processing and on the result page. This identifier additionally allows users to share the result with other researchers.

The current implementation offers several graphical and tabular outputs in addition to the processed sequence data. The Coverage versus Identity plot shows the number of matching reads for different coverage and identity threshold values. The coverage plots show where the metatranscriptomic sequences aligned to the rRNA reference sequences and provide an easy way to check for possible bias in the alignment or the rRNA-removal prior to sequencing. The coverage data for each database sequence is available for download. The taxonomic classifications of rRNA-like sequences are presented as bar charts for each selected database. The summary report includes information about the input data, selected databases and thresholds, and rRNA-like sequence classifications by database, domain and phyla.

4 BRIEF SURVEY OF ALTERNATIVE PROGRAMS

There are different applications that can identify rRNA-like sequences in metatranscriptomic datasets. The command line program meta_rna (Huang *et al*., 2009) is written in Python and identifies rRNA sequences based on HMMs using the HMMER package (Eddy, 2009). Another program based on HMMER is rRNASelector (Lee *et al*., 2011), which is written in Java and can only be used through its graphical interface. The web-based MG-RAST (Meyer *et al*., 2008) uses the BLASTn program, identifying rRNA-like sequences based on sequence similarity. The HMM-based programs currently allow identification of bacterial and archaeal rRNAs. The sequence similarity-based programs make it easy to assign sequences to taxonomic groups.

5 CONCLUSION

riboPicker allows scientists to efficiently remove rRNA-like sequences from their metatranscriptomic datasets prior to downstream analysis. The web interface is simple and user-friendly, and the stand-alone version allows offline analysis and integration into existing data processing pipelines. The tool provides a computational resource able to handle the amount of data that next-generation sequencers are capable of generating and can place the process more within reach of the average research lab.

ACKNOWLEDGEMENT

We thank Matthew Haynes and Ramy Aziz for comments and suggestions. We thank the HMP DACC for making reference genomes data from the NIH Human Microbiome Project publicly available.

Funding: National Science Foundation Advances in Bioinformatics grant (DBI 0850356 to R.E.).

Conflict of Interest: none declared.

REFERENCES

Aziz,R.K. *et al*. (2008) The RAST Server: Rapid Annotations using Subsystems Technology. *BMC Genomics*, **9**, 75.

Cole,J.R. *et al*. (2009) The Ribosomal Database Project: improved alignments and new tools for rRNA analysis. *Nucleic Acids Res.*, **37**, D141–D145.

DeSantis,T.Z. *et al*. (2006) Greengenes, a Chimera-Checked 16S rRNA Gene Database and Workbench Compatible with ARB. *Appl. Environ. Microbiol*., **72**, 5069–5072.

Eddy,S.R. (2009) A new generation of homology search tools based on probabilistic inference. *Genome Inform*., **23**, 205–211.

Gardner,P.P. *et al*. (2011) Rfam: Wikipedia, clans and the "decimal" release. *Nucleic Acids Res*., **39**, D141–D145.

He,S. *et al*. (2010) Validation of two ribosomal RNA removal methods for microbial metatranscriptomics. *Nat. Methods*, 7, 807–812.

Huang,Y. *et al*. (2009) Identification of ribosomal RNA genes in metagenomic fragments. *Bioinformatics*, **25**, 1338–1340.

Li,H. and Durbin, R. (2010) Fast and accurate long-read alignment with Burrows-wheeler transform. *Bioinformatics*., **26**, 589–595.

Lee,J.-H. *et al*. (2011) rRNASelector: a computer program for selecting ribosomal RNA encoding sequences from metagenomic and metatranscriptomic shotgun libraries. *J. Microbiol*., **49**, 689–691.

Meyer,F. *et al*. (2008) The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics*, **9**, 386.

Pruesse,E. *et al*. (2007) SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Res*., **35**, 7188–7196.

Sayers,E.W. *et al*. (2011) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res*., **39**, D38–D51.

Schmieder,R. and Edwards,R. (2011) Quality control and preprocessing of metagenomic datasets. *Bioinformatics*, **27**, 863–864.

Stewart,F.J. *et al*. (2010) Development and quantitative analyses of a universal rRNA-subtraction protocol for microbial metatranscriptomics. *ISME J.*, **4**, 896–907.

The NIH HMP Working Group *et al*. (2009) The NIH Human Microbiome Project. *Genome Res*., **19**, 2317–2323.

Tripp,H.J. *et al*. (2011) Misannotations of rRNA can now generate 90% false positive protein matches in metatranscriptomic studies. *Nucleic Acids Res.*, **39**, 8792–8802.