Bioinformatics, 31(13), 2015, 2199–2201
doi: 10.1093/bioinformatics/btv106
Advance Access Publication Date: 19 February 2015
Applications Note

OXFORD

Sequence analysis

Annocript: a flexible pipeline for the annotation of transcriptomes able to identify putative long noncoding RNAs

Francesco Musacchia[1,†], Swaraj Basu[1,†], Giuseppe Petrosino[1], Marco Salvemini[2] and Remo Sanges[1,*]

[1]Biology and Evolution of Marine Organisms, Stazione Zoologica "Anton Dohrn", Villa Comunale, 80121, Naples, Italy and [2]Department of Biology, University of Naples Federico II, Via Mezzocannone 8, 80134, Naples, Italy

*To whom correspondence should be addressed.
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.
Associate Editor: John Hancock

Received on October 26, 2014; revised on January 22, 2015; accepted on February 11, 2015

Abstract

Summary: The eukaryotic transcriptome is composed of thousands of coding and long non-coding RNAs (lncRNAs). However, we lack a software platform to identify both RNA classes in a given transcriptome. Here we introduce Annocript, a pipeline that combines the annotation of protein coding transcripts with the prediction of putative lncRNAs in whole transcriptomes. It downloads and indexes the needed databases, runs the analysis and produces human readable and standard outputs together with summary statistics of the whole analysis.
Availability and implementation: Annocript is distributed under the GNU General Public License (version 3 or later) and is freely available at https://github.com/frankMusacchia/Annocript.
Contact: remo.sanges@szn.it

1 Introduction

RNA-Seq has effectively portrayed the transcriptional complexity in eukaryotes demonstrating the widespread transcription of lncRNAs in a diverse group of organisms. However, annotation of *de novo* generated transcriptomes, remains a complex task requiring efficient management of large datasets. Current tools for annotation identify coding sequences by similarity searches (Conesa *et al.*, 2005; Koski *et al.*, 2005; Philipp *et al.*, 2012; Schmid and Blaxter, 2008) while lncRNAs are typically predicted relying on sequence composition, lack of conservation and lack of ORFs (Kong *et al.*, 2007; Lin *et al.*, 2011; Liu *et al.*, 2006; Wang *et al.*, 2013). To fulfill the need of a single platform, we developed Annocript, a pipeline linking together these processes. It uses multiple programs to annotate sequences based on similarity and composition. It is capable of assigning protein names, domains, metabolic pathways (Morgat *et al.*, 2012), Enzyme classes (Bairoch, 2000), GO terms (Ashburner *et al.*, 2000) and other annotations. Further, it can identify putative lncRNAs based upon lack of any protein/domain similarity, lack of long ORFs and high non-coding potential. By default, a transcript is classified as lncRNA if it satisfies all the following conditions: (i) length $\geq$200 nucleotides, (ii) lack of similarity with any protein, domain and other short ncRNA from Rfam and rRNAs, (iii) longest open reading frame (ORF) $<$ 100 amino acids and (iv) non-coding potential score $\geq$0.95.

2 Implementation

Annocript will annotate any FASTA containing a transcriptome given as input. There is no limit in the number of sequences it can analyze. The pipeline runs on Linux systems and requires the presence of (tested versions): Perl (5.10), BioPerl (1.6) (Stajich *et al.*, 2002), Python (2.7.3), R (3.1.0) and MySQL (5.5). A single configuration file allows the user to alter programs parameters or run only specific analysis steps. Annocript performs the following analysis steps.

2.1 Database creation

This step downloads the sequence and annotation databases required for similarity searches: UniRef or TrEMBL coupled with SwissProt (UniProt Consortium, 2013), Rfam (Burge *et al.*, 2012) and Conserved Domains Database (CDD) (Marchler-Bauer *et al.*, 2013). The UniRef/TrEMBL and SwissProt FASTA headers are used to build a MySQL database containing information on taxonomy, protein id, name and description. The downloaded FASTA are indexed for BLAST searches (Camacho *et al.*, 2009). Annocript also downloads ID mappings to associate UniProt and Pfam IDs to Enzymes (Bairoch 2000), GO and pathways into the MySQL database. The database construction requires a few hours on high memory machines. A pure Perl hash table is created to build relations between the tables before to store them into the database. This usage of Perl hash tables is a memory-intensive task that permits to build the database quickly (Table 1). Nevertheless, the pipeline can be run on low-memory machines with as little as 4 GB RAM using a different approach and longer computational time. By default Annocript requires 20 GB to build the hash table and this would be sufficient to build a database using UniRef90 + SwissProt. In our tests the peak amount of RAM used by the hash when loading the entire UniProt knowledge-base (TrEMBL + SwissProt, August 2014), is 44 GB. The database creation step is needed only for the first run of Annocript or when the user needs to update it.

2.2 Programs execution

This step is responsible for the execution of the programs. It performs the following similarity searches: BLASTX against TrEMBL/UniRef and SwissProt, RPSBLAST against CDD profiles, BLASTN against Rfam and rRNAs. By default, Annocript uses speed-optimized BLASTX and BLASTP parameters and a custom parallelization of RPSBLAST to achieve a faster execution. The dna2pep (Wernersson *et al.*, 2006) and Portrait softwares (Arrial *et al.*, 2009) are used to identify, respectively, the longest ORF and the non-coding potential of each query sequence.

2.3 Results parsing and statistics

This step integrates all the results generated in previous steps building a comprehensive tab delimited text file and other standard outputs. The pipeline collects the best-hit and related annotations from each search. The output table contains specific columns for every search executed. In this way, the user can decide to choose a specific classification for cases of uncertainty based on scores or lengths or any other criteria. The columns of the file represent assigned proteins, domains, GO terms, Enzymes, pathways, short and ribosomal RNAs, longest ORF size and non-coding potential. Information about the score, position, coverages and combination of domains are also provided. The table provides a binary classification for each transcript to be a protein coding or a lncRNA. Annocript also generates FASTA files with the predicted lncRNAs, protein sequences from the longest ORFs and GFF3 files of complete BLAST outputs. Finally, a summary of the annotation is given as a HTML-formatted document. Examples of whole transcriptome annotations made by Annocript can be downloaded from http://bit.ly/15vnALW.

Table 1. Time required to build the internal MySQL database

	L	M	S	XS
TrEMBL + SwissProt	9 h	24 h	NA	NA
UniRef90 + SwissProt	5 h	11 h	40 h	120 h

Note: Time to create the MySQL database from TrEMBL + SwissProt or UniRef90 + SwissProt (April 2014) using machines with different number of cores and RAM: 24 cores/98GB (L), 24 cores/24GB (M), 8 cores/8GB (S), 2 cores/4GB (XS).

3 Optimization

To reduce the computational time, Annocript uses the following parameters in BLASTX and BLASTP searches: threshold = 18 and word_size = 4 (Korf 2003). We show that this customization does not affect the results on two sample datasets: transcripts from a species present in UniProt and from a species not present in UniProt at the time of analysis. The two datasets contain, respectively, 1500 coding sequences from *Homo sapiens* and 1500 transcripts from *Acipenser naccarii* (Vidotto *et al.*, 2013). Annocript was executed for each species with BLAST default and Annocript optimized BLAST parameters. The results for human were identical while for the sturgeon, using optimized parameters, out of 899 transcripts annotated in the BLAST default run, 891 (99%) were identical, 3 (0.003%) were missing and 5 matched on different, but homolog hits. At a negligible cost in sensitivity, the computational time was reduced more than 5-folds (Table 2).

4 Discussion

Annocript is a pipeline for the annotation of transcriptomes. The pipeline is optimized to run quickly without bargaining on accuracy. It is the first platform with the ability to annotate coding genes along with the prediction of putative lncRNAs without the need for a reference genome or comparative supporting data.

Acknowledgements

The authors acknowledge the laboratories of Luigia Santella and Jong Tai Chun, Graziano Fiorito, Mariella Ferrante and Gabriele Procaccini for their feedback. S.B. has been funded by the Stazione Zoologica under the ARC PhD program of The Open University, UK.

Funding

RITMARE Flagship Project, Progetto Premiale StarTrEgg.

Conflict of Interest: none declared.

Table 2. Time required to run searches with default and optimized parameters

Organism	BLASTX SP def	BLASTX SP cust	BLASTX Uf def	BLASTX Uf cust	RPSBLAST CDD def	RPSBLAST CDD cust
Human	50.2 min	10.9 min	33.4 h	5.2 h	9.4 h	67.5 min
Sturgeon	12.8 min	2.07 min	8.2 h	54.3 min	2.7 h	14.4 min

Note: Analysis performed on human coding transcripts (avg. length 4200 bp) and *A. naccarii* transcripts (avg. length 1200 bp) using 24 cores (2.67 GHz) and 24 GB RAM and default BLAST parameters (def) or Annocript-optimized parameters (cust).

References

Arrial,R.T. *et al*. (2009) Screening noncoding RNAs in transcriptomes from neglected species using PORTRAIT: case study of the pathogenic fungus *Paracoccidioides brasiliensis*. *BMC Bioinformatics*, **10**, 239.

Ashburner,M. *et al*. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, **25**, 25–29.

Bairoch,A. (2000) The ENZYME database in 2000. *Nucleic Acids Res.*, **28**, 304–305.

Burge,S.W. *et al*. (2012) Rfam 11.0: 10 years of RNA families. *Nucleic Acids Res.*, **41**, D226–D232.

Camacho,C. *et al*. (2009) BLAST+: architecture and applications. *BMC Bioinformatics*, **10**, 421.

Conesa,A. *et al*. (2005) Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics*, **21**, 3674–3676.

Kong,L. *et al*. (2007) CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine. *Nucleic Acids Res.*, **35**, W345–W349.

Korf,I. (2003) Serial BLAST searching. *Bioinformatics*, **19**, 1492–1496.

Koski,L.B. *et al*. (2005) AutoFACT: an automatic functional annotation and classification tool. *BMC Bioinformatics*, **6**, 151.

Lin,M.F. *et al*. (2011) PhyloCSF: a comparative genomics method to distinguish protein coding and noncoding regions. *Bioinformatics*, **27**, i275–i282.

Liu,J. *et al*. (2006) Distinguishing protein-coding from noncoding RNAs through support vector machines. *PLoS Genet.*, **2**, e29.

Marchler-Bauer,A. *et al*. (2013) CDD: conserved domains and protein three-dimensional structure. *Nucleic Acids Res.*, **41**, D348–D352.

Morgat,A. *et al*. (2012) UniPathway: a resource for the exploration and annotation of metabolic pathways. *Nucleic Acids Res.*, **40**, D761–D769.

Philipp,E.E.R. *et al*. (2012) The transcriptome analysis and comparison explorer–T-ACE: a platform-independent, graphical tool to process large RNAseq datasets of non-model organisms. *Bioinformatics*, **28**, 777–783.

Schmid,R. and Blaxter,M.L. (2008) annot8r: GO, EC and KEGG annotation of EST datasets. *BMC Bioinformatics*, **9**, 180.

Stajich,J.E. *et al*. (2002) The Bioperl Toolkit: Perl Modules for the Life Sciences. *Genome Res.*, **12**, 1611–1618.

UniProt Consortium. (2013) Update on activities at the Universal Protein Resource (UniProt) in 2013. *Nucleic Acids Res.*, **41**, D43–D47.

Vidotto,M. *et al*. (2013) Transcriptome sequencing and de novo annotation of the critically endangered Adriatic sturgeon. *BMC Genomics*, **14**, 407.

Wang,L. *et al*. (2013) CPAT: Coding-potential assessment tool using an alignment-free logistic regression model. *Nucleic Acids Res.*, **41**, e74.

Wernersson,R. (2006) Virtual Ribosome–a comprehensive DNA translation tool with support for integration of sequence feature annotation. *Nucleic Acids Res.*, **34**, W385–W388.