Methods 63 (2013) 60–65

Contents lists available at SciVerse ScienceDirect

Methods

journal homepage: www.elsevier.com/locate/ymeth

Detection of non-coding RNA in bacteria and archaea using the DETR'PROK Galaxy pipeline ☆

Claire Toffano-Nioche [a], Yufei Luo [b], Claire Kuchly [a], Claire Wallon [a], Delphine Steinbach [b], Matthias Zytnicki [b], Annick Jacq [a], Daniel Gautheret [a,*]

[a] Université Paris-Sud, Institut de Génétique et Microbiologie, CNRS UMR 8621, Orsay F-91405, France
[b] URGI, INRA, Versailles F-78026, France

ARTICLE INFO

Article history:
Available online 25 June 2013

Keywords:
Non-coding RNA
Bacteria
Archaea
Small RNA
Antisense RNA
Untranslated regions

ABSTRACT

RNA-seq experiments are now routinely used for the large scale sequencing of transcripts. In bacteria or archaea, such deep sequencing experiments typically produce 10–50 million fragments that cover most of the genome, including intergenic regions. In this context, the precise delineation of the non-coding elements is challenging. Non-coding elements include untranslated regions (UTRs) of mRNAs, independent small RNA genes (sRNAs) and transcripts produced from the antisense strand of genes (asRNA). Here we present a computational pipeline (DETR'PROK: detection of ncRNAs in prokaryotes) based on the Galaxy framework that takes as input a mapping of deep sequencing reads and performs successive steps of clustering, comparison with existing annotation and identification of transcribed non-coding fragments classified into putative 5′ UTRs, sRNAs and asRNAs. We provide a step-by-step description of the protocol using real-life example data sets from *Vibrio splendidus* and *Escherichia coli*.

© 2013 The Authors. Published by Elsevier Inc. All rights reserved.

1. Introduction

The discovery and annotation of non-coding RNA genes (ncRNAs) in bacteria and archaea is a long term endeavor that began in the 1990s and is far from being finished today even for well studied model genomes. The functions of ncRNAs include major life-sustaining tasks such as translation (rRNAs, tRNAs) and ancient housekeeping roles (guide RNAs, RNAse P, SRP RNA). The majority of newly discovered ncRNAs in bacteria and archaea however belong to three specific classes: small RNAs (sRNAs) are generally involved in trans-regulation activities mediated by the binding of target mRNAs [1], long 5′ UTRs contain cis-acting RNAs such as riboswitches [2], T-boxes [3] and other classes of translation or transcription attenuators [4], while cis-encoded antisense RNAs (asRNAs) are RNAs produced from the opposite strand of coding or non-coding genes, some of which may contribute to RNA processing [5].

Early ncRNA detection in bacteria relied strongly on bioinformatics and, more specifically, comparative genomics that identifies non-coding regions with unexpected conservation among closely related species. However, since the advent of next generation sequencing (NGS) and its application to high throughput sequencing of transcripts (RNA-seq), biologists have turned to this method to accelerate the pace of ncRNA discovery. A major benefit of RNA-seq over bioinformatics is that it can identify non-conserved RNAs and antisense RNAs. RNA-seq screens for ncRNAs have now been performed on most model genomes and have contributed to a large increase of the lists of ncRNAs in these species. However the use of RNA-seq for ncRNA identification is only beginning as the number of genomes to be analyzed is at least three orders of magnitude larger than the number already analyzed. Moreover, even in model genomes, RNA-seq experiments will need to be carried out in multiple different conditions in order to enable the discovery of RNAs expressed under specific growth conditions.

As an illustration of how ncRNA annotation lags behind DNA sequencing, consider that most current publicly available genome annotations have no ncRNA annotated other than tRNAs and rRNAs. For instance, while the *Escherichia coli* K12 MG1655 strain NCBI annotation file has a solid set of 63 annotated ncRNAs, this is an exception, as even closely related *E. coli* strains have little or no ncRNAs annotated. Species such as *Salmonella enterica*, *Bacillus subtilis* or *Staphylococcus aureus*, in which hundreds of ncRNA have been published and stored in public databases such as RFAM [6], have their annotation files devoid of any ncRNA gene (except for tRNAs and rRNAs). Furthermore, since automated annotation pipelines are not able to identify actual transcription starts and stops, all genes annotated in bacterial genomes are limited to their

☆ This is an open-access article distributed under the terms of the Creative Commons Attribution-NonCommercial-No Derivative Works License, which permits non-commercial use, distribution, and reproduction in any medium, provided the original author and source are credited.

* Corresponding author.
E-mail address: daniel.gautheret@u-psud.fr (D. Gautheret).

1046-2023/$ - see front matter © 2013 The Authors. Published by Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.ymeth.2013.06.003

coding part and are thus devoid of 5′ and 3′ UTRs. Finally, cis-encoded antisense RNAs are completely absent from current annotations.

Processing RNA-seq data to identify new ncRNAs and UTRs in a sequenced genome is a complex task. First, RNA-seq produces short reads that must be mapped onto the genome. Mapped reads must then be clustered and compared to existing annotation, so that clusters can be associated to known genes or deemed "intergenic" or "antisense". Because RNA-seq reads do not cover transcribed regions homogeneously, defining ncRNA elements requires several parameters specifying minimum coverage, distance between elements, etc. In this article we present a protocol that performs this analysis using a set of publicly available tools, plugged together using the Galaxy framework [7–10]. We present here the main steps of the workflow, explain user-defined parameters and show examples of applications to bacterial genomes.

Running the workflow requires that the user can access a Galaxy instance with the DETR'PROK workflow installed. We assume users have access to such an instance and know the general usage of Galaxy commands. Basic Galaxy tutorials are available at http://usegalaxy.org/.

2. The DETR'PROK workflow

A simplified view of the DETR'PROK workflow is shown in Fig. 1. The program requires as input an alignment file in BAM format and an annotation file in GFF format. The BAM alignment file can be produced by any next-generation sequencing read mapper, such as Bowtie [11] or BWA [12]. As our workflow is highly dependent on read orientation to differentiate sense from antisense transcripts, the initial sequencing must be produced using an oriented RNA-seq library preparation protocol. We recommend to run the mapping program in the "unique match" mode and/or to discard reads corresponding to highly expressed loci such as rRNA genes. This produces lighter BAM files thus leading to faster runs and file transfers. The annotation file is important, as UTR extensions and novel independent sRNAs will be defined relative to this prior annotation. If the annotation file contains ncRNA annotations already (with "ncRNA" in the 3rd column), then these ncRNAs will not be predicted again, unless users require their annotations are ignored. Note that, as said earlier, most NCBI bacterial genome files do not contain ncRNA annotation so far.

Starting from the BAM and GFF file, the workflow clusters overlapping reads and compares clusters to previous annotations to produce "extended annotations", that is tentative transcription units containing annotated ncRNAs or CDS, extended using the RNA-seq clusters. Extended annotations are further analyzed to produce lists of sRNAs, antisense RNAs, UTRs and operon spacers, in the form of GFF annotation files. These steps are further detailed below.

3. Workflow steps and parameters

The DETR'PROK workflow is composed of more than 40 steps. Most steps are based on the S-MART toolbox [13], a set of tools that allows a convenient handling of RNA-seq mapping results to

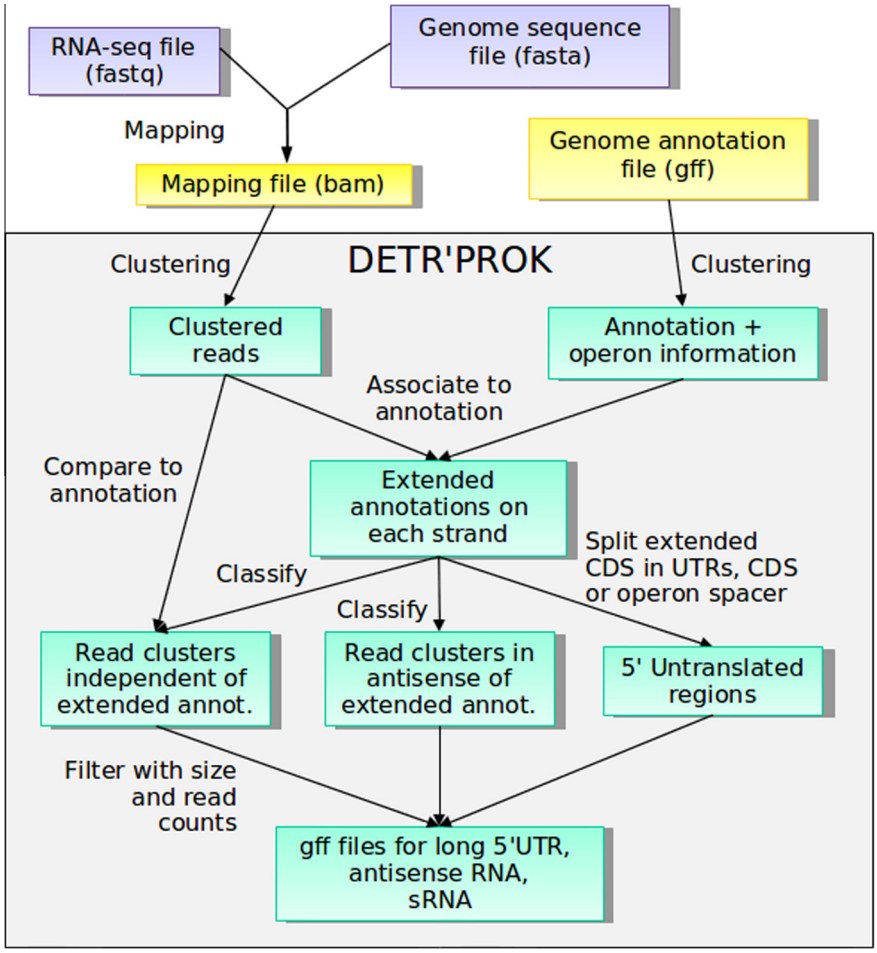

Fig. 1. Simplified view of the DETR'PROK workflow. Color code is as follows. Yellow: workflow input data; green: workflow steps; purple: other data.

perform read clustering, intersection with genomic segments and production of new annotation files. Instead of combining the 40+ steps into a single "black box" procedure, we opted for exposing the independent steps in an open workflow, in accordance with the general philosophy of the Galaxy system. In this way, our workflow can easily be taken apart, adapted or improved by any Galaxy-literate user.

Most steps in DETR'PROK use fixed parameters and hence do not require any user input. However, user input is required at 14 steps. The corresponding parameters are presented in Fig. 2 and Table 1. There are fewer parameters than steps because the same parameter is used at several steps (this is a normal constraint in the Galaxy framework). The first parameter (features_list) is used to specify which initial annotation in the input GFF file should be preserved. By not including ncRNAs or tRNAs at this step, these elements are ignored and possibly re-predicted by the workflow. Note that some GFF annotation files may use other names to describe ncRNAs, such as "misc_RNA". Most other parameters cover two types of items: distance constraints between elements to separate or merge elements, and filters to reject ncRNAs based on their sizes or number of mapped reads. Importantly, the default DETR'PROK pipeline requires asRNAs to be covered completely by their antisense transcript (which can be coding or non-coding). We opted for this stringent definition of antisense because in the compact bacterial and archaeal genomes a large fraction of transcripts overlap transcripts from the opposite DNA strand and allowing sRNAs to overlap their opposite transcripts partially may thus yield high levels of false positives. This default behavior can be changed by manually editing the corresponding parameters in the workflow, as explained in section "Changing a fixed parameter".

For each parameter, a suggested value or range of values is displayed in the comments of the Galaxy workflow. In our experience with several bacterial and archaeal RNA-seq analyzes, there is no set of predefined values that will work satisfyingly for all (or even most) species and RNA-seq data. Two major factors weighing on parameter settings are gene density and sequencing depth. As gene density increases, the gaps (RNA_gap parameter) separating elements should be made shorter to avoid merging different elements together. The "op_gap" parameter defining the maximum internal operon spacers is also correlated to gene density. It can be tentatively defined based on the distribution of intergenic distances in the whole genome: multicistronic genes tend to create a peak of short intergenic distances that can be used to set the "op_gap" parameter just above this peak (Fig. S1). As sequencing depth increases, so does the likelihood of spurious reads in antisense or intergenic regions, therefore requiring that coverage filters (sRNA_min_reads, asRNA_min_reads, UTR5_min_reads) are set higher. The complete pipeline runs fast enough (about 2 h on a standard Linux server with an average sized bacterial genome and 5 M sequencing reads) that parameter values can be improved by trial and error.

Once the complete workflow is executed, distinct GFF annotation files are produced for long 5′ UTRs, sRNAs and asRNAs (file names ending respectively with long_5UTR_list.gff, sRNA_list.gff and asRNA_list.gff), as well as for interesting byproducts of the workflow: 5′ and 3′ extensions of all initial CDS and ncRNAs, and intra-operon spacers (file names ending respectively with 5_extension_list.gff, 3_extension_list.gff and operon_spacer_list.gff). Each element is described by one line of the corresponding GFF file, providing information about chromosome, position, strand. A tag named "nbOverlappingReads" provides the total number of reads

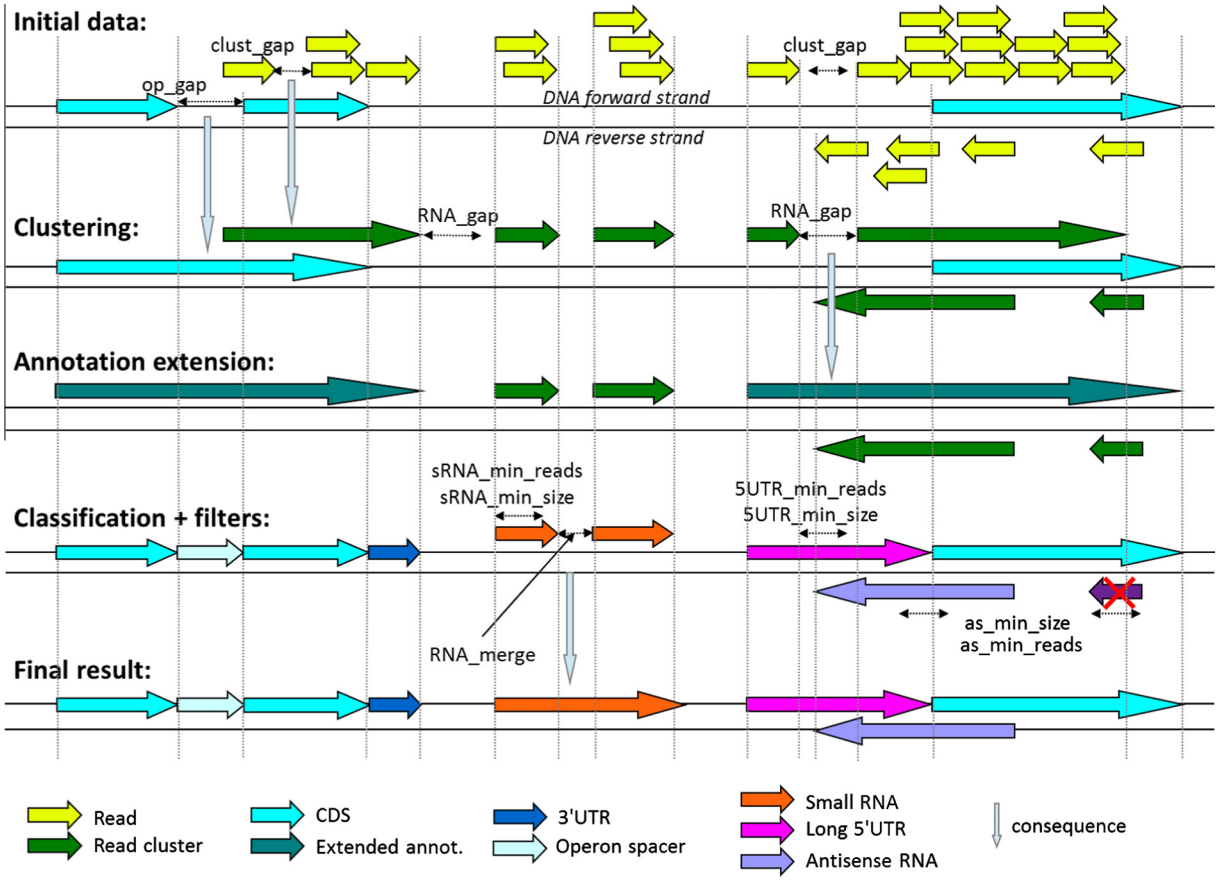

Fig. 2. A graphical description of user-defined parameters in the DETR'PROK workflow. Parameters are shown at their respective workflow step. Steps are presented from top to bottom in sequential order.

Table 1
Workflow parameters: description, name, name of step in workflow and example settings.

Constraint	Parameter	Step name in workflow	Settings used for tests
List of features to use as initial annotation	Features_list	Clean transcript file	rRNA, tRNA, CDS (*V. splendidus*), ncRNA, rRNA, tRNA, CDS (*E. coli*)
Maximal intergenic distance within operon	Op_gap	Clusterize	150 (*V. splendidus*), 30 (*E. coli*)
Maximal gap between reads in a cluster	Clust_gap	Clusterize	20
Maximal gap between a cluster and a CDS for definition of extended annotation	RNA_gap	Compare overlapping small query, Clusterize	25
Maximal distance to merge independent RNA candidates	RNA_merge	Clusterize	50
Minimal number of reads for sRNA, asRNA, 5′ UTR	sRNA_min_reads, asRNA_min_reads 5UTR_min_reads	Select by tag	12 22 10
Minimal size for sRNA, asRNA, 5′ UTR, respectively	sRNA_min_size, asRNA_min_size, 5UTR_min_size	Restrict from size	50

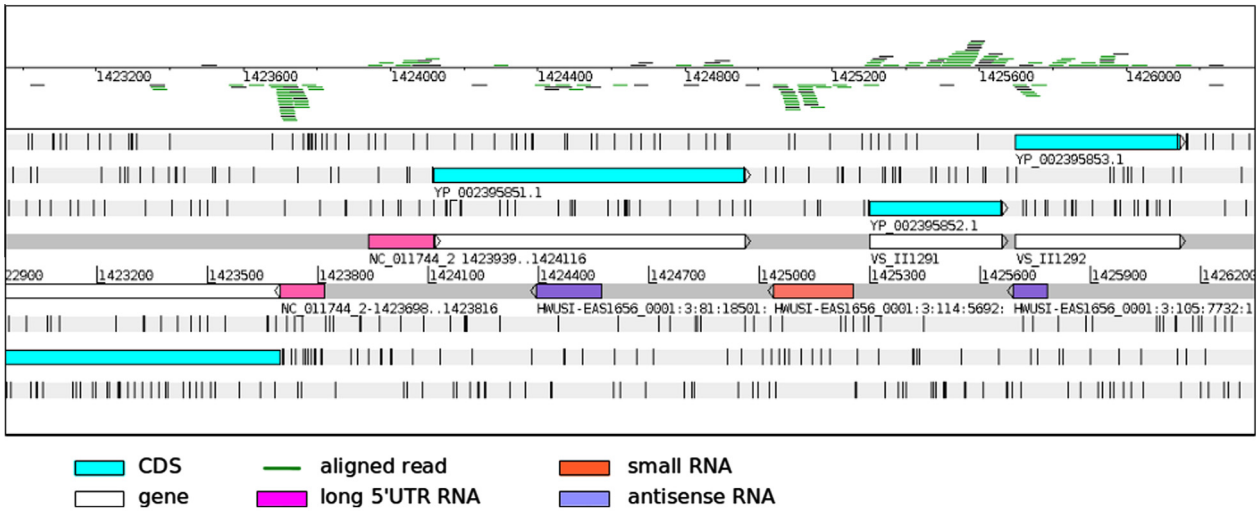

Fig. 3. Novel annotated ncRNAs in the *V. splendidus* genome, visualized using the Artemis browser. Top panel: aligned RNA-seq reads. Position above or below the line indicate read orientations. Central panel: annotations. Top three and bottom three lines in central panel indicate open reading frames (blue) and stop codons (vertical bars). The rightmost antisense RNA is considered as such as it is strictly included in the extended annotation composed by the two CDS above grouped as an operon by DETR'PROK (extended annotation is not shown).

covering the element. The GFF files can be examined either directly as text using the Galaxy file viewer or saved on the user's computer and visualized using a genome viewer such as Artemis [14,15]. Note that genome viewers often require indexed BAM files (.bai format) for displaying NGS reads along the genome sequence. A bai file is automatically created when adding a BAM file to a dataset into Galaxy data libraries.

Once a parameter set is deemed satisfying for a given genome and sequencing depth, the workflow can be saved with these parameters by extracting the workflow from the history (select "Extract Workflow" from the History menu on top of the right panel). Alternatively, parameters can be hard-coded into the workflow using the workflow editor (Workflow menu, select a workflow, select Edit, then click on a Workflow step and, for a given parameter, change "to be set at runtime" with "set in advance"). In this way, the pipeline can be run again over similar genomes and RNA-seq datasets with a single click.

4. A test run using a *Vibrio splendidus* RNA-seq dataset

We reanalyzed a RNA-seq dataset obtained for the oyster pathogen *V. splendidus*, a gamma proteobacteria that contains two chromosomes of size 3.3 Mb and 1.7 Mb. Original sequencing conditions were single-end Illumina GA-IIX sequencing, 38 nt reads, 28 M reads overall. This transcriptome was analyzed using an earlier version of our pipeline followed by manual curation [16]. Here we analyzed again the same dataset using the latest DETR'PROK workflow and no manual curation.

We obtained the *V. splendidus* GFF annotation files directly from the NCBI FTP site using the URL:

ftp://ftp.ncbi.nlm.nih.gov/genomes/Bacteria/Vibrio_splendidus_LGP32_uid59353/

The annotation files NC_011753.gff and NC_011744.gff were imported into Galaxy by selecting "Get Data" – "Upload File from your computer". The mapping procedure that generates BAM files can be run within Galaxy using a program such as Bowtie. Inputs are the Fasta format genome sequence and the original Fastq file produced by the sequencing machine (here SRA accession SRX272401). This step is not part of the DETR'PROK workflow and we assume it has been performed independently. This resulted in 4.5 M reads mapping at unique positions [16]. The corresponding BAM files are available at:

http://rna.igmors.u-psud.fr/suppl_data/

The parameter settings for the DETR'PROK workflow are shown in Table 1. These were refined empirically to approximate those used in our previous analysis (see Fig. 2 in Ref. [16]).

The previous workflow and manual curation had identified 250 sRNAs, 471 long 5′ UTR and 73 asRNAs [16]. The new DETR'PROK workflow identifies 242 sRNAs, 519 long 5′ UTR and 86 asRNAs. Numbers of identified ncRNAs are thus quite similar between the

Table 2
Recall of known *E. coli* sRNAs from a single directional RNA-seq experiment performed in normal growth conditions (SRA accession SRR358747).

Total annotated sRNAs (NCBI NC_000913.2)	63 (including RNAse P RNA, 6S RNA, and SRP RNA)
Total recalled by DETR'PROK	35 (56%)
# Recalled as sRNA	34
# Recalled but misclassified	1 (SRP RNA called as long 5′ UTR)
# Missed and expressed	20 (3 embedded in CDS, 2 in operon spacers -6S RNA-, and 15 due to number of reads below threshold)
# Missed and silent	8
# Additional sRNAs	312

two procedures. Differences in numbers are imputable to several factors such as a version change of the *V. splendidus* annotation file that includes additional sRNAs, a slightly different definition of antisense between the two protocols and the manual curation applied to the earlier version. We loaded the three ncRNA GFF output files for visualization in Artemis. A color encoding in the GFF file permits an easy differentiation of the RNA types. A subset of RNA predictions is shown in Fig. 3 in their genomic environment.

5. Recall of trusted ncRNAs from *E. coli*

We questioned the ability of a single RNA-seq experiment analyzed with the DETR'PROK workflow to recall the ncRNA complement in an organism where most ncRNAs are supposedly identified. Currently, *E. coli* has the largest annotated ncRNA collection, thanks to multiple screens performed over the past ten years under a variety of growth conditions. The current *E. coli* K12 MG 1655 annotation (accession NC_000913.2) contains 63 sRNAs. Antisense RNAs are absent and long 5′ UTRs are too unevenly annotated (only represented by a dozen leader peptides) to be tested. We retrieved RNA-seq reads from the SRA accession SRR358747, which contains a directional RNA library from *E. coli* K-12 MG1655 grown in LB to log phase, sequenced (single end, 35-nt reads) with an Illumina GA-IIX sequencer [17]. We aligned reads to the *E. coli* K-12 MG1655 genome using a standard mapping protocol (Suppl. Methods) and sampled two millions (about 20%) of the uniquely aligned reads for input to the DET'RPROK workflow. The workflow identified 346 sRNAs, 153 asRNAs and 720 long 5′ UTRs overall. Table 2 presents the recall statistics of annotated sRNAs. Thirty five of the 63 annotated sRNAs (56%) were correctly recalled. Most of the missed sRNAs were lost because they were not expressed at all (8 sRNAs) or their expression was below threshold (15 sRNAs). One sRNA was misclassified as a long 5′ UTR, two were found as operon spacers and three could not be detected because they were embedded in a CDS of the initial annotation. Note that although the overall recall rate was only 56%, this only reflects the inherent limitation of a single experiment in detecting all sRNAs, as the workflow did not miss any significantly expressed sRNA. Conversely, the large number of additional sRNAs identified by the workflow (312) is not a computational artifact. These RNAs are all supported by 12 or more reads. Most likely, a large fraction of these RNAs is known already, as the RFAM database [6] contains about a hundred more *E. coli* ncRNAs than annotated in the current NCBI genome sequence.

6. Conclusion and further developments

We presented a workflow for the annotation of non-coding RNAs in archaeal and bacterial genomes. This workflow requires some user input in order to take genome density and sequencing depth into account. However, we showed that the Galaxy framework enables parameters to be fixed permanently for a faster launch procedure. In our final output files, we singled out long 5′ UTRs, sRNAs and asRNAs because these elements constitute a current focus of the RNA community as they are most likely to harbor functional, regulatory RNAs. However non-coding RNA is also present in the form of 3′ UTRs and short 5′ UTRs. These elements are interesting because they redefine the transcription units of coding genes, which is important for transcript quantification and for the study of RNA–RNA interactions. Furthermore, archaeal RNAs may have long and possibly functional 3′ UTRs [18,19]. Users interested in 5′ and 3′ UTRs of any size can retrieve the corresponding GFF files from the workflow.

7. Workflow installation

The workflow should be run on a local Galaxy instance or one hosted on a distant server. Refer to the tutorial (http://wiki.galaxy-project.org/Admin/Get_Galaxy) for creation of a Galaxy instance. As the workflow has several steps running in parallel, the Galaxy instance should use a database management system supporting concurential access such as postgreSQL, instead of the default SQLite system. The DETR'PROK pipeline is obtained from the Galaxy main tool shed (http://toolshed.g2.bx.psu.edu/), or my experiment (http://www.myexperiment.org), searching by name "detrprok_wf", and should be installed by the Galaxy instance administrator. DETR'PROK requires installation of the S-MART toolbox [13] and the "detrprok_scripts", both available in the Galaxy main tool shed.

8. Changing a fixed parameter in a Galaxy workflow

Fixed parameters are not accessible to users at run time. However, expert users can change fixed parameters, such as the stringency of antisense overlap, by manually editing the workflow steps. To this aim, users should select Workflow in the top Galaxy menu, select the "Detrprok" workflow, select Edit and find the box corresponding to the workflow step to be changed. At this stage, each parameter for this step can be edited by entering values in the corresponding boxes. Changes are saved by clicking on the top right gear-wheel in the central editing panel. All parameter values can be modified in the same way. The parameter defining the stringency of antisense definition is named "minOverlap" and is used twice, in the last two "compare overlapping small query" steps of the workflow. The "minOverlap" value is the minimal overlap between an asRNA and its antisense element, expressed as a percentage of the asRNA length (range: 0–100).

9. File type glossary

- BAM file: a BAM file is a binary file containing information about the alignment of RNA-seq reads onto a reference genome sequence. BAM is the binary version of the SAM (Sequence Alignment/Map) format. BAM to SAM conversions (and reverse) can be performed easily using the Galaxy framework [20].
- BAI file: a BAI file is an index file and is paired with a BAM file.
- GFF file: a GFF (General Feature Format) file is a tabular text file providing annotation information for a genome. Each line describes a feature such as a gene, exon, intron, coding sequence, ncRNA, tRNA, etc. For each feature, a GFF file provides information such as chromosome name, start, end, strand, direction, etc.
- FASTQ file: A FASTQ file is a text file storing both DNA sequences and their corresponding quality scores. Most

high throughput sequencers now produce FASTQ files as output.
- SRA file: a SRA (Short Read Archive) file is an archive file storing both DNA sequences and their corresponding metadata. The FASTQ sequence file can be extracted from the SRA archive using fastq-dump, a part of NCBI's SRA toolkit.

Acknowledgements

This work was funded in part by Agence Nationale pour la Recherche (grant ANR-2010-BLAN-1602-01) 'Duplex-omics', and by the GIS-IBISA. We thank the eBio Bioinformatics platform for technical support.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.ymeth.2013.06.003.

References

[1] K.M. Wassarman, F. Repoila, C. Rosenow, G. Storz, S. Gottesman, Genes Dev. 15 (2001) 1637–1651.
[2] W.C. Winkler, R.R. Breaker, Annu. Rev. Microbiol. 59 (2005) 487–517.
[3] F.J. Grundy, T.M. Henkin, Cell 74 (1993) 475–482.
[4] C. Yanofsky, RNA 13 (2007) 1141–1154.
[5] I. Lasa, A. Toledo-Arana, A. Dobin, M. Villanueva, I. Ruiz de los Mozos, M. Vergara-Irigara, et al., Proc. Natl. Acad. Sci. USA 108 (2011) 20172–20177.
[6] S.W. Burge, J. Daub, R. Eberhardt, J. Tate, L. Barquist, E.P. Nawrocki, et al., Nucl. Acids Res. 41 (2013) D226–D232.
[7] J. Goecks, A. Nekrutenko, J. Taylor and The Galaxy Team, Genome Biol. 11 (2010) R86.
[8] D. Blankenberg, G. Von Kuster, N. Coraor, G. Ananda, R. Lazarus, M. Mangan, A. Nekrutenko, J. Taylor, Curr. Protoc. Mol. Biol. 89 (2010) 19.10.1–19.10.21.
[9] B. Giardine, C. Riemer, R.C. Hardison, R. Burhans, L. Elnitski, P. Shah, Y. Zhang, D. Blankenberg, I. Albert, J. Taylor, W. Miller, W.J. Kent, A. Nekrutenko, Genome Res. 15 (2005) 1451–1455.
[10] R. Lazarus, A. Kaspi, M. Ziemann and The Galaxy Team, Bioinformatics 28 (2012) 3139–3140.
[11] B. Langmead, C. Trapnell, M. Pop, S.L. Salzberg, Genome Biol. 10 (2009) R25.
[12] H. Li, R. Durbin, Bioinformatics 25 (2009) 1754–1760.
[13] M. Zytnicki, H. Quesneville, PLoS One 6 (2011) e25988.
[14] K. Rutherford, J. Parkhill, J. Crook, T. Horsnell, P. Rice, Bioinformatics 16 (2000) 944–945.
[15] T. Carver, S.R. Harris, M. Berriman, J. Parkhill, J.A. McQuillan, Bioinformatics 28 (2012) 464–469.
[16] C. Toffano-Nioche, RNA 18 (2012) 2201–2219.
[17] R. Raghavan, D.B. Sloan, H. Ochman, MBio 3 (2012). e00156-12.
[18] M. Brenneis, O. Hering, C. Lange, J. Soppa, PLoS Genet. 3 (2007) e229.
[19] J. Straub, M. Brenneis, A. Jellen-Ritter, R. Heyer, J. Soppa, A. Marchfelder, RNA Biol. 6 (2009) 281–292.
[20] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, Bioinformatics 25 (2009) 2078–2079.