CircParser: a novel streamlined pipeline for circular RNA structure and host gene prediction in non-model organisms

Artem Nedoluzhko [Corresp., Equal first author, 1], **Fedor Sharko** [Equal first author, 2, 3], **Md. Golam Rbbani** [1], **Anton Teslyuk** [4], **Ioannis Konstantinidis** [1], **Jorge MO Fernandes** [Corresp. 1]

[1] Faculty of Biosciences and Aquaculture, Nord University, Bodø, Bodø, Norway

[2] Genome Center, National Research Centre "Kurchatov Institute", Moscow, Russia, Russia

[3] Institute of Bioengineering, Research Center of Biotechnology, Moscow, Russia, Russia

[4] National Research Centre "Kurchatov Institute", Moscow, Russia, Russia

Corresponding Authors: Artem Nedoluzhko, Jorge MO Fernandes
Email address: artem.nedoluzhko@nord.no, jorge.m.fernandes@nord.no

Circular RNAs (circRNAs) are long noncoding RNAs which play a significant role in various biological processes, including embryonic development and stress responses. These regulatory molecules can modulate microRNA activity and be involved in different molecular pathways as indirect regulators of gene expression. Thousands of circRNAs have been described in diverse taxa due to the recent advances in high throughput sequencing technologies, which led to a huge variety of total RNA sequencing being publicly available. A number of circRNA *de novo* and host gene prediction tools are available to date, but their ability to accurately predict circRNA host genes is limited in the case of low-quality genome assemblies or annotations. Here, we present CircParser, a simple and fast Unix/Linux pipeline that uses the outputs from the most common circular RNAs *in silico* prediction tools (CIRI, CIRI2, CircExplorer2, find_circ, and circFinder) to annotate circular RNAs, assigning presumable host genes from local or public databases such as National Center for Biotechnology Information (NCBI). Also this pipeline can discriminate circular RNAs based on their structural components (exonic, intronic, exon-intronic or intergenic) using genome annotation file .

CircParser: a novel streamlined pipeline for circular RNA structure and host gene prediction in non-model organisms

Artem Nedoluzhko[1,*], Fedor Sharko[2,3,*], Md. Golam Rbbani[1], Anton Teslyuk[3], Ioannis Konstantinidis[1], and Jorge M. O. Fernandes[1]

[1] Faculty of Biosciences and Aquaculture, Nord University, Bodø, Nordland county, Norway
[2] Institute of Bioengineering, Research Center of Biotechnology of the Russian Academy of Sciences, Moscow, Russia
[3] Complex of NBICS Technologies, National Research Centre "Kurchatov Institute", Moscow, Russia

Corresponding Author:
Artem Nedoluzhko[1]
Universitetsalléen 11, Bodø, Nordland county, 8049, Norway
Email address: artem.nedoluzhko@nord.no
Fedor Sharko[2,3]
Kurchatov sq. 1, Moscow, 123182, Russia
Email address: fedosic@gmail.com
Jorge M. O. Fernandes[1]
Universitetsalléen 11, Bodø, Nordland county, 8049, Norway
Email address: jorge.m.fernandes@nord.no

Abstract

Circular RNAs (circRNAs) are long noncoding RNAs which play a significant role in various biological processes, including embryonic development and stress responses. These regulatory molecules can modulate microRNA activity and be involved in different molecular pathways as indirect regulators of gene expression. Thousands of circRNAs have been described in diverse taxa due to the recent advances in high throughput sequencing technologies, which led to a huge variety of total RNA sequencing being publicly available. A number of circRNA *de novo* and host gene prediction tools are available to date, but their ability to accurately predict circRNA host genes is limited in the case of low-quality genome assemblies or annotations.
Here, we present CircParser, a simple and fast Unix/Linux pipeline that uses the outputs from the most common circular RNAs *in silico* prediction tools (CIRI, CIRI2, CircExplorer2, find_circ, and circFinder) to annotate circular RNAs, assigning presumable host genes from local or public databases such as National Center for Biotechnology Information (NCBI). Also this pipeline can discriminate circular RNAs based on their structural components (exonic, intronic, exon-intronic or intergenic) using genome annotation file.

Introduction

De novo genome sequencing has become a routine procedure, due to a decrease in sequencing costs, diversification of high-throughput sequencing platforms and improvement of bioinformatic tools (Ekblom and Wolf, 2014). However, the quality of non-model species genome assemblies and, as a result, their annotations are often of unsatisfactory quality, because of (1) repetitive sequences, including transposons, and short sequence repeats (SSRs); (2) gene and genome duplications; (3) single-nucleotide polymorphisms (SNPs) and genome rearrangements (Lien, et al., 2016; Negrisolo, et al., 2010; Rodriguez and Arkhipova, 2018; Yahav and Privman, 2019).
CircRNAs are relatively poorly studied members of the non-coding RNA family. These unique single-stranded molecules are generated through back-splicing of pre-mRNAs in a wide range of eukaryotic and prokaryotic taxa (Danan, et al., 2012; Holdt, et al., 2018), and even viruses (Huang, et al., 2019). CircRNAs play a significant role in the regulation of the molecular pathways not only through modulating of microRNA and protein activity, but also by the affecting transcription or splicing (Holdt, et al., 2018).
These regulatory molecules have been known for decades, but the development of high-throughput DNA analysis methods lead to a rapid increase in the number of studies related to these type of non-coding RNAs. This, in turn, resulted in a requirement for additional circRNA prediction tools. The miARma-Seq (Andres-Leon and Rojas, 2019) with CIRI predictor (Gao, et al., 2015), circRNA_finder (Westholm, et al., 2014), find_circ (Memczak, et al., 2013), CIRCexplorer2 (Zhang, et al., 2016), and other tools are very popular today for prediction of circRNAs sequences based on transcriptomic data (Hansen, et al., 2016; Szabo and Salzman, 2016), despite significant output differences. Several circRNA predictors (CIRI, CIRI2, and CircExplorer2) can use genome annotation files for host gene prediction but they are definitely useful only for well-annotated genomes, and even, such as CircView (Feng, et al., 2018) or circMeta (Chen, et al., 2019), have been designed for them.
Here we describe CircParser, a novel, easy to use and Unix/Linux pipeline for circular RNAs host gene prediction using the blastn program and the freely available bedtools software (Quinlan and Hall, 2010). CircParser can be also implemented as a part of pipelines for *de novo* prediction of circular RNA because of its versatile output files. CircParser is most useful for circRNA host gene prediction analysis in whole transcriptomic datasets for low-quality assembled as well as poorly annotated genomes. It sorts and joins overlapped circular RNAs sequences and predicts host gene name for overrepresented circRNAs, while identifying their structural components. We demonstrate the prediction capacity of CircParser on a recently published transcriptomic data set from the wild and domesticated females of Nile tilapia (Oreochromis niloticus) fast muscle (Konstantinidis et al., under review) using the five most popular circRNAs *in silico* prediction tools – CIRI, CIRI2, CircExplorer2, find_circ, and circFinder.

Materials & Methods

The results of Illumina sequencing of twelve ribosomal RNA depleted RNA-seq libraries reads have been downloaded from Gene Expression Omnibus (accession number GSE135811). The DNA reads were filtered by quality (phred > 20) and library adapters were trimmed using Cutadapt software (version 1.12) (Marcel, 2011). The Nile tilapia reference genome (ASM185804v2) and its gene-annotation (ref_O_niloticus_UMD_NMBU_top_level.gff3) were used in the following analysis.

CircRNA prediction was performed for each ribosomal RNA depleted RNA-seq library using the circRNA *in silico* prediction tools i) CIRI (Gao, et al., 2015) that is linked to miARma-Seq pipeline (Andres-Leon and Rojas, 2019), ii) CIRI2 (Gao, et al., 2018), iii) CircExplorer2 (Zhang, et al., 2016), iv) find_circ (Memczak, et al., 2013), and v) circFinder (Westholm, et al., 2014). Prediction output files from all libraries were converted separately to coordinate file format. After sorting, these coordinate files (from different prediction algorithms, but for each library) were merged using bedtools multiinner (Quinlan and Hall, 2010) to determine a joint prediction output from CIRI, CIRI2, CircExplorer2, find_circ, and circFinder (see Supplementary Table S1).

We developed CircParser, as a streamlined pipeline, which makes use output files from the most popular circRNAs *in silico* predictors. CircParser only works under Linux/Unix system. The parameters for CircParser are presented in Table 1.

Usage: perl CircParser.pl [-h] -b INPUT_FILE –genome REF_GENOME

[Table 1]

CircParser can merge overlapped circRNAs coordinates from circRNAs predictor outputs using bedtools merge (Quinlan and Hall, 2010) at the first stage of the pipeline; this ensures that they are related to the same host gene and creates separate coordinates files (bed file) with overlapped circRNAs coordinates. In addition, it is optionally possible to merge circRNA without overlapping coordinates but located in the contiguous genome locus using the special option. The separate coordinate files (bed file) are converted to fasta files using bedtools getfasta (Quinlan and Hall, 2010). Finally, CircParser uses fasta files for host gene prediction using a NCBI database (the longest stage of pipeline) for circRNAs (Figure 1A). CircParser works by default with the NCBI online database, but it can optionally use a custom database or a pre-compiled NCBI database installed locally.

[Figure 1]

CircParser can also discriminate circular RNAs by their structural components: exonic, intronic, exon-intronic or intergenic using genome annotation gff/gff3 file (-a parameter). In this case, the user should avoid circRNAs coordinate merging (using --np parameter) during the pipeline implementation for correct results (Figure 1B).

Usage: perl CircParser.pl -np -b INPUT_FILE –genome REF_GENOME -a GENOME.gff
However, poor quality of annotation file can lead to errors in the circRNAs structure analysis.
The Perl implementation of CircParser is available at https://github.com/SharkoTools/CircParser

Results and discussion

We applied CircParser to twelve merged coordinate files that contained information about joint coordinates for circRNAs predicted using CircExplorer2, miARma-Seq (with CIRI predictor), CIRI2, find_circ, and circFinder. The five different algorithms predicted on average ~131 (CircExplorer2); ~501 (CIRI); ~706 (CIRI2); ~257 (find_circ), and ~398 (circFinder) circRNAs per sample, with an insignificant overlap ~37 circRNAs (Figure 2; Supplementary Table S1), similarly to previously published comparisons (Hansen, 2018; Hansen, et al., 2016).

[Figure 2]

To access the host gene of circular RNAs and to reduce of false-positive rates only overlapping circRNAs (Figure 2) were used in CircParser. This pipeline allows the elimination of non-informative outputs (e.g contains only chromosome/contig name, number of uncharacterized loci, or name of BAC clone, and etc.), while keeping more the relevant blast results and retrieving the likely host gene name for the circular RNAs; in the case of impossibility to find identical sequences in the database, this tool mark these sequence as NOT ASSIGNED).
The CircParser results also allow to determine the number of circRNA types from one host gene and their minimum and maximum size in base pairs (bp). We showed that our algorithm detected presumable host gene names for the vast majority of predicted circRNAs. Moreover, most of them were related to muscle functions (e.g. *calcium/calmodulin-dependent protein kinase*, *troponin T3*, *myocyte-specific enhancer factor 2C*, and others), and immune-related genes (*MHC class IA antigen*), which consistently found among different individuals (Supplementary Table S2), despite the relatively low coverage (for circRNAs analysis) of used sequencing data for the circRNA analysis (Mahmoudi and Cairns, 2019). The example of circRNA structure analysis for CIRI, CIRI2, CircExplorer2, find_circ, and circFinder outputs are presented in Supplementary Table S3.
We conclude that CircParser represents a fast and reproducible workflow that enables researchers to predict the host genes for circular RNAs, even in non-model organisms with poorly annotated genome assemblies.

Acknowledgements

We would like to acknowledge Jorge Galindo-Villegas from Nord University (Norway) and Tomas B. Hansen from the Aarhus University (Denmark) for their valuable advices.

References

1. Andres-Leon, E. and Rojas, A.M. miARma-Seq, a comprehensive pipeline for the simultaneous study and integration of miRNA and mRNA expression data. Methods 2019;152:31-40.
2. Chen, L., et al. circMeta: a unified computational framework for genomic feature annotation and differential expression analysis of circular RNAs. Bioinformatics 2019.
3. Danan, M., et al. Transcriptome-wide discovery of circular RNAs in Archaea. Nucleic Acids Res 2012;40(7):3131-3142.
4. Ekblom, R. and Wolf, J.B. A field guide to whole-genome sequencing, assembly and annotation. Evol Appl 2014;7(9):1026-1042.
5. Feng, J., et al. CircView: a visualization and exploration tool for circular RNAs. Brief Bioinform 2018;19(6):1310-1316.
6. Gao, Y., Wang, J. and Zhao, F. CIRI: an efficient and unbiased algorithm for *de novo* circular RNA identification. Genome Biol 2015;16:4.
7. Gao, Y., Zhang, J. and Zhao, F. Circular RNA identification based on multiple seed matching. Brief Bioinform 2018;19(5):803-810.
8. Hansen, T.B. Improved circRNA Identification by Combining Prediction Algorithms. Front Cell Dev Biol 2018;6:20.
9. Hansen, T.B., et al. Comparison of circular RNA prediction tools. Nucleic Acids Res 2016;44(6):e58.
10. Holdt, L.M., Kohlmaier, A. and Teupser, D. Molecular roles and function of circular RNAs in eukaryotic cells. Cell Mol Life Sci 2018;75(6):1071-1098.
11. Huang, J.T., et al. Identification of virus-encoded circular RNA. Virology 2019;529:144-151.
12. Lien, S., et al. The Atlantic salmon genome provides insights into rediploidization. Nature 2016;533(7602):200-205.
13. Mahmoudi, E. and Cairns, M.J. Circular RNAs are temporospatially regulated throughout development and ageing in the rat. Sci Rep 2019;9(1):2564.
14. Marcel, M. Cutadapt Removes Adapter Sequences From High-Throughput Sequencing Reads. EMBnet.journal 2011;17(1):10-12.
15. Memczak, S., et al. Circular RNAs are a large class of animal RNAs with regulatory potency. Nature 2013;495(7441):333-338.
16. Negrisolo, E., et al. Different phylogenomic approaches to resolve the evolutionary relationships among model fish species. Mol Biol Evol 2010;27(12):2757-2774.
17. Quinlan, A.R. and Hall, I.M. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 2010;26(6):841-842.
18. Rodriguez, F. and Arkhipova, I.R. Transposable elements and polyploid evolution in animals. Curr Opin Genet Dev 2018;49:115-123.
19. Szabo, L. and Salzman, J. Detecting circular RNAs: bioinformatic and experimental challenges. Nat Rev Genet 2016;17(11):679-692.

20. Westholm, J.O., et al. Genome-wide analysis of drosophila circular RNAs reveals their structural and sequence properties and age-dependent neural accumulation. Cell Rep 2014;9(5):1966-1980.
21. Yahav, T. and Privman, E. A comparative analysis of methods for *de novo* assembly of hymenopteran genomes using either haploid or diploid samples. Sci Rep 2019;9(1):6480.
22. Zhang, X.O., et al. Diverse alternative back-splicing and alternative splicing landscape of circular RNAs.

Table 1(on next page)

Table 1. CircParser.pl usage. Required and optional parameters

Table 1. CircParser.pl usage. Required and optional parameters

1 Table 1. CircParser.pl usage. Required and optional parameters

Parameter	Parameter description
-h, --help	Show this help message and exit
-b	CircRNA input file (required)
-g, --genome	Reference genome file (required)
-t, --tax	NCBI TaxID (optional)
-a	Genome annotation file, gff/gff3 file (optional)
--np	Prohibition for coordinate merging (optional)
-c, --ciri	Input circRNA from CIRI|CIRI2 *in silico* predictors, (default: input from CircExplorer2, find_circ, circFinder, and BED files)
--threads	Number of threads (CPUs) for BLAST search (optional)
-v, --version	Current CircParser version

2

Figure 1

Figure 1

An overview of the CircParser pipeline: **Figure 1A**: The pipeline that includes merging of the circRNAs with overlapping genome coordinates and presents the number of different circRNAs originating from one host gene. **Figure 1B**: Annotation of each predicted circRNA.

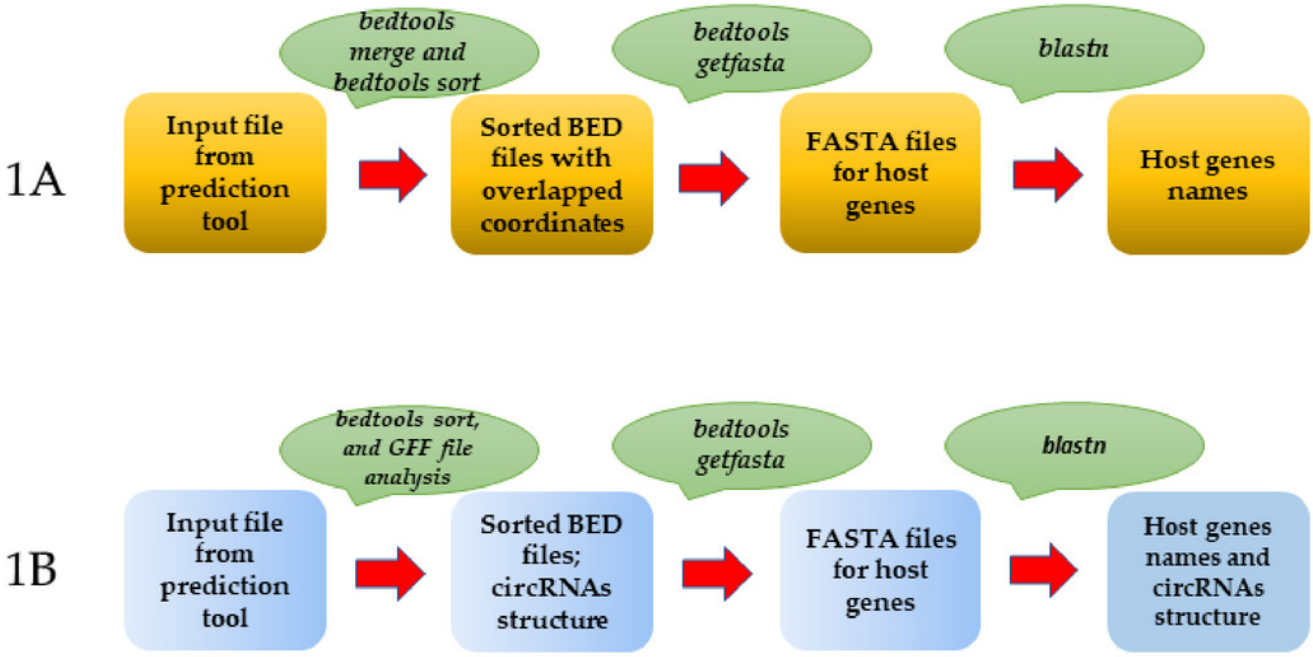

Figure 2

Figure 2

Number of circular RNAs, that have been predicted by CIRI, CIRI2, CircExplorer2, find_circ, circFinder, and that are common between all prediction algorithms.