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** [2], **Ioannis Konstantinidis** [1], **Jorge MO Fernandes** [Corresp. 1]

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

[2] Complex of NBICS Technologies, National Research Centre "Kurchatov Institute", Moscow, Russia

[3] Institute of Bioengineering, Research Center of Biotechnology of the Russian Academy of Sciences, 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 are 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.

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

4 Artem Nedoluzhko[1,*], Fedor Sharko[2,3,*], Md. Golam Rbbani[1], Anton Teslyuk[3], Ioannis
5 Konstantinidis[1], and Jorge M. O. Fernandes[1]
6
7 [1] Faculty of Biosciences and Aquaculture, Nord University, Bodø, Nordland county,
8 Norway
9 [2] Institute of Bioengineering, Research Center of Biotechnology of the Russian Academy
10 of Sciences, Moscow, Russia
11 [3] Complex of NBICS Technologies, National Research Centre "Kurchatov Institute",
12 Moscow, Russia
13
14 Corresponding Author:
15 Artem Nedoluzhko[1]
16 Universitetsalléen 11, Bodø, Nordland county, 8049, Norway
17 Email address: artem.nedoluzhko@nord.no
18 Jorge M. O. Fernandes[1]
19 Universitetsalléen 11, Bodø, Nordland county, 8049, Norway
20 Email address: jorge.m.fernandes@nord.no
21

22 **Abstract**

23 Circular RNAs (circRNAs) are long noncoding RNAs which play a significant role in
24 various biological processes, including embryonic development and stress responses.
25 These regulatory molecules can modulate microRNA activity and are involved in
26 different molecular pathways as indirect regulators of gene expression. Thousands of
27 circRNAs have been described in diverse taxa due to the recent advances in high
28 throughput sequencing technologies, which led to a huge variety of total RNA
29 sequencing being publicly available. A number of circRNA *de novo* and host gene
30 prediction tools are available to date, but their ability to accurately predict circRNA host
31 genes is limited in the case of low-quality genome assemblies or annotations.
32 Here, we present CircParser, a simple and fast Unix/Linux pipeline that uses the
33 outputs from the most common circular RNAs *in silico* prediction tools (CIRI, CIRI2,
34 CircExplorer2, find_circ, and circFinder) to annotate circular RNAs, assigning

35 presumable host genes from local or public databases such as National Center for

36 Biotechnology Information (NCBI). Also, this pipeline can discriminate circular RNAs

37 based on their structural components (exonic, intronic, exon-intronic or intergenic)

38 using genome annotation file.

39

40 **Introduction**

41 *De novo* genome sequencing has become a routine procedure, due to a decrease in

42 sequencing costs, diversification of high-throughput sequencing platforms and

43 improvement of bioinformatic tools (Ekblom & Wolf 2014). However, the quality of

44 non-model species genome assemblies and, as a result, their annotations are often of

45 unsatisfactory quality, because of (1) repetitive sequences, including transposons, and

46 short sequence repeats (SSRs); (2) gene and genome duplications; (3) single-nucleotide

47 polymorphisms (SNPs) and genome rearrangements (Lien et al. 2016; Negrisolo et al.

48 2010; Rodriguez & Arkhipova 2018; Yahav & Privman 2019).

49 CircRNAs are relatively poorly studied members of the non-coding RNA family. These

50 unique single-stranded molecules are generated through back-splicing of pre-mRNAs

51 in a wide range of eukaryotic and prokaryotic taxa (Danan et al. 2012; Holdt et al. 2018),

52 and even viruses (Huang et al. 2019). CircRNAs play a significant role in the regulation

53 of the molecular pathways not only through modulating of microRNA and protein

54 activity, but also by the affecting transcription or splicing (Holdt et al. 2018).

55 These regulatory molecules have been known for decades, but the development of high-

56 throughput DNA analysis methods lead to a rapid increase in the number of studies

57 related to these type of non-coding RNAs. This, in turn, resulted in a requirement for

58 additional circRNA prediction tools. The miARma-Seq (Andres-Leon & Rojas 2019)

59 with CIRI predictor (Gao et al. 2015), circRNA_finder (Westholm et al. 2014), find_circ

60 (Memczak et al. 2013), CIRCexplorer2 (Zhang et al. 2016), and other tools are very

61 popular today for prediction of circRNAs sequences based on transcriptomic data

62 (Hansen et al. 2016; Szabo & Salzman 2016), despite significant output differences.

63 Several circRNA predictors (CIRI, CIRI2, and CircExplorer2) can use genome

64 annotation files for host gene prediction but they are definitely useful only for well-

65 annotated genomes, and even, such as CircView (Feng et al. 2018) or circMeta (Chen et

66 al. 2019), have been designed specifically for them.

67 Here we describe CircParser, a novel and easy to use Unix/Linux pipeline for prediction

68 of host gene circular RNAs using the blastn program and the freely available bedtools

69 software (Quinlan & Hall 2010). CircParser can be also implemented as a part of
70 pipelines for *de novo* prediction of circular RNA because of its versatile output files.
71 CircParser is most useful for circRNA host gene prediction analysis in whole
72 transcriptomic datasets for low-quality assembled, as well as poorly annotated
73 genomes. It sorts and joins overlapped circular RNAs sequences and predicts host gene
74 name for overrepresented circRNAs, while identifying their structural components. We
75 demonstrate the prediction capacity of CircParser on a recently published
76 transcriptomic data set from the wild and domesticated females of Nile tilapia
77 (*Oreochromis niloticus*) fast muscle (Konstantinidis et al., under review) using the five
78 most popular circRNAs *in silico* prediction tools – CIRI, CIRI2, CircExplorer2, find_circ,
79 and circFinder.
80
81

82 **Materials & Methods**
83 The results of Illumina sequencing of twelve ribosomal RNA depleted RNA-seq
84 libraries reads have been downloaded from Gene Expression Omnibus (accession
85 number GSE135811). The DNA reads were filtered by quality (phred > 20) and library
86 adapters were trimmed using Cutadapt software (version 1.12) (Marcel 2011). The Nile
87 tilapia reference genome (ASM185804v2) and its gene-annotation
88 (ref_O_niloticus_UMD_NMBU_top_level.gff3) were used in the following analysis.
89 CircRNA prediction was performed for each ribosomal RNA depleted RNA-seq library
90 using the circRNA *in silico* prediction tools i) CIRI (Gao et al. 2015) that is linked to
91 miARma-Seq pipeline (Andres-Leon & Rojas 2019), ii) CIRI2 (Gao et al. 2018), iii)
92 CircExplorer2 (Zhang et al. 2016), iv) find_circ (Memczak et al. 2013), and v) circFinder
93 (Westholm et al. 2014). Prediction output files from all libraries were converted
94 separately to coordinate file format. After sorting, these coordinate files (from different
95 prediction algorithms, but for each library) were merged using bedtools multiinter
96 (Quinlan & Hall 2010) to determine a joint prediction output from CIRI, CIRI2,
97 CircExplorer2, find_circ, and circFinder (see Supplementary Table S1).
98 We developed CircParser, as a streamlined pipeline, which makes use output files from
99 the most popular circRNAs *in silico* predictors. CircParser only works under Linux/Unix
100 system. The parameters for CircParser are presented in Table 1.
101 *Usage: perl CircParser.pl [-h] -b INPUT_FILE –genome REF_GENOME*
102

103   [Table 1]

104

105   CircParser can merge overlapped circRNAs coordinates from circRNAs predictor

106   outputs using bedtools merge (Quinlan & Hall 2010) at the first stage of the pipeline;

107   this ensures that they are related to the same host gene and creates separate coordinates

108   files (bed file) with overlapped circRNAs coordinates. In addition, it is optionally

109   possible to merge circRNA without overlapping coordinates but located in the

110   contiguous genome locus using the special option.

111   The separate coordinate files (bed file) are converted to fasta files using bedtools

112   getfasta (Quinlan & Hall 2010). Finally, CircParser uses fasta files for host gene

113   prediction using a NCBI database (the longest stage of pipeline) for circRNAs (Figure

114   1A). CircParser works by default with the NCBI online database, but it can optionally

115   use a custom database or a pre-compiled NCBI database installed locally. CircParser

116   includes the following blast parameters, which are necessary for host gene prediction,

117   and assigns sequences to the respective circRNA: *-perc_identity* 90; *-max_target_seqs* 1000;

118   *-max_hsps* 1; the maximum number of aligned sequences to keep is 1000; the minimum

119   percent identity of matches to report is 90 %. CircParser also filters out non-informative

120   blast results, such as "uncharacterized", "clone", "linkage group" and others from the

121   output table.

122

123   [Figure 1]

124

125   CircParser can also discriminate circular RNAs by their structural components: exonic,

126   intronic, exon-intronic or intergenic using genome annotation gff/gff3 file (-a

127   parameter). In this case, the user should avoid circRNAs coordinate merging (using --

128   np parameter) during the pipeline implementation for correct results (Figure 1B).

129   *Usage: perl CircParser.pl -np -b INPUT_FILE –genome REF_GENOME -a GENOME.gff*

130   However, poor quality of annotation file can lead to errors in the circRNAs structure

131   analysis.

132   The Perl implementation of CircParser is available at

133   https://github.com/SharkoTools/CircParser

134

135   **Results**

136   We applied CircParser to twelve merged coordinate files that contained information
137   about joint coordinates for circRNAs predicted using CircExplorer2, miARma-Seq (with
138   CIRI predictor), CIRI2, find_circ, and circFinder. The five different algorithms predicted
139   on average ~131 (CircExplorer2); ~501 (CIRI); ~706 (CIRI2); ~257 (find_circ), and ~398
140   (circFinder) circRNAs per sample, with an insignificant overlap ~37 circRNAs (Figure
141   2; Supplementary Table S1), similarly to previously published comparisons (Hansen
142   2018; Hansen et al. 2016).

143

144   [Figure 2]

145

146   To access the host gene of circular RNAs and to reduce false-positive rates, only
147   overlapping circRNAs (Figure 2) were used in CircParser. This pipeline allows the
148   elimination of non-informative outputs (e.g contains only chromosome/contig name,
149   number of uncharacterized loci, or name of BAC clone, and etc.), while keeping more
150   the relevant blast results and retrieving the likely host gene name for the circular RNAs;
151   in the case of impossibility to find identical sequences in the database, this tool mark
152   these sequence as NOT ASSIGNED).

153

154   **Discussion**
155   The CircParser results also allow to determine the number of circRNA types from one
156   host gene and their minimum and maximum size in base pairs (bp). We showed that
157   our algorithm detected presumable host gene names for the vast majority of predicted
158   circRNAs. Moreover, most of them were related to muscle functions (e.g.
159   *calcium/calmodulin-dependent protein kinase*, *troponin T3*, *myocyte-specific enhancer factor 2C*,
160   and others), and immune-related genes (*MHC class IA antigen*), which were consistently
161   found among different individuals (Supplementary Table S2), despite the relatively low
162   coverage for circRNAs analysis of the sequencing data used (Mahmoudi & Cairns 2019).
163   An example of circRNA structure analysis for CIRI, CIRI2, CircExplorer2, find_circ, and
164   circFinder outputs are presented in Supplementary Table S3.
165   To estimate the capacity of our pipeline we compared a number of host genes
166   that were predicted by CircExplorer2 and CircParser (CircExplorer2 outputs were used
167   as input files) for the same *O. niloticus* fast muscle datasets used earlier. As a result,
168   CircParser shows greater efficiency for Nile tilapia, improving the number of predicted
169   host genes up to two-fold (Figure 3).

170

171   [Figure 3]

172

173   Another equally important aspect of CircParser concerned the accuracy of this pipeline.

174   The most well-annotated reference genome of zebrafish (assembly GRCz11) and

175   zebrafish muscle transcriptomic dataset (ERR145655) were used for accuracy

176   estimation, i.e. the agreement between the annotation file and CircParser output. We

177   showed that in this case, CircParser host gene prediction was confirmed in 82.4% cases.

178

179   **Conclusions**

180   Thus, we conclude that CircParser represents a reproducible workflow that enables

181   researchers to effectively predict the host genes for circular RNAs, even in non-model

182   organisms with poorly annotated genome assemblies.

183

184

185   **Acknowledgements**

186   We would like to acknowledge Jorge Galindo-Villegas from Nord University (Norway)

187   and Tomas B. Hansen from the Aarhus University (Denmark) for their valuable advice.

188

189

190

191   **References**

192

193   Andres-Leon E, and Rojas AM. 2019. miARma-Seq, a comprehensive pipeline for the
194   simultaneous study and integration of miRNA and mRNA expression data.
195   *Methods* 152:31-40. 10.1016/j.ymeth.2018.09.002
196   Chen L, Wang F, Bruggeman EC, Li C, and Yao B. 2019. circMeta: a unified
197   computational framework for genomic feature annotation and differential
198   expression analysis of circular RNAs. *Bioinformatics*.
199   10.1093/bioinformatics/btz606
200   Danan M, Schwartz S, Edelheit S, and Sorek R. 2012. Transcriptome-wide discovery of
201   circular RNAs in Archaea. *Nucleic Acids Res* 40:3131-3142. 10.1093/nar/gkr1009
202   Ekblom R, and Wolf JB. 2014. A field guide to whole-genome sequencing, assembly and
203   annotation. *Evol Appl* 7:1026-1042. 10.1111/eva.12178

204 Feng J, Xiang Y, Xia S, Liu H, Wang J, Ozguc FM, Lei L, Kong R, Diao L, He C, and Han
205   L. 2018. CircView: a visualization and exploration tool for circular RNAs. *Brief*
206   *Bioinform* 19:1310-1316. 10.1093/bib/bbx070
207 Gao Y, Wang J, and Zhao F. 2015. CIRI: an efficient and unbiased algorithm for de novo
208   circular RNA identification. *Genome Biol* 16:4. 10.1186/s13059-014-0571-3
209 Gao Y, Zhang J, and Zhao F. 2018. Circular RNA identification based on multiple seed
210   matching. *Brief Bioinform* 19:803-810. 10.1093/bib/bbx014
211 Hansen TB. 2018. Improved circRNA Identification by Combining Prediction
212   Algorithms. *Front Cell Dev Biol* 6:20. 10.3389/fcell.2018.00020
213 Hansen TB, Veno MT, Damgaard CK, and Kjems J. 2016. Comparison of circular RNA
214   prediction tools. *Nucleic Acids Res* 44:e58. 10.1093/nar/gkv1458
215 Holdt LM, Kohlmaier A, and Teupser D. 2018. Molecular roles and function of circular
216   RNAs in eukaryotic cells. *Cell Mol Life Sci* 75:1071-1098. 10.1007/s00018-017-2688-
217   5
218 Huang JT, Chen JN, Gong LP, Bi YH, Liang J, Zhou L, He D, and Shao CK. 2019.
219   Identification of virus-encoded circular RNA. *Virology* 529:144-151.
220   10.1016/j.virol.2019.01.014
221 Lien S, Koop BF, Sandve SR, Miller JR, Kent MP, Nome T, Hvidsten TR, Leong JS,
222   Minkley DR, Zimin A, Grammes F, Grove H, Gjuvsland A, Walenz B,
223   Hermansen RA, von Schalburg K, Rondeau EB, Di Genova A, Samy JK, Olav Vik
224   J, Vigeland MD, Caler L, Grimholt U, Jentoft S, Vage DI, de Jong P, Moen T,
225   Baranski M, Palti Y, Smith DR, Yorke JA, Nederbragt AJ, Tooming-Klunderud A,
226   Jakobsen KS, Jiang X, Fan D, Hu Y, Liberles DA, Vidal R, Iturra P, Jones SJ,
227   Jonassen I, Maass A, Omholt SW, and Davidson WS. 2016. The Atlantic salmon
228   genome provides insights into rediploidization. *Nature* 533:200-205.
229   10.1038/nature17164
230 Mahmoudi E, and Cairns MJ. 2019. Circular RNAs are temporospatially regulated
231   throughout development and ageing in the rat. *Sci Rep* 9:2564. 10.1038/s41598-
232   019-38860-9
233 Marcel M. 2011. Cutadapt Removes Adapter Sequences From High-Throughput
234   Sequencing Reads. *EMBnetjournal* 17:10-12. 10.14806/ej.17.1.200.
235 Memczak S, Jens M, Elefsinioti A, Torti F, Krueger J, Rybak A, Maier L, Mackowiak SD,
236   Gregersen LH, Munschauer M, Loewer A, Ziebold U, Landthaler M, Kocks C, le
237   Noble F, and Rajewsky N. 2013. Circular RNAs are a large class of animal RNAs
238   with regulatory potency. *Nature* 495:333-338. 10.1038/nature11928
239 Negrisolo E, Kuhl H, Forcato C, Vitulo N, Reinhardt R, Patarnello T, and Bargelloni L.
240   2010. Different phylogenomic approaches to resolve the evolutionary
241   relationships among model fish species. *Mol Biol Evol* 27:2757-2774.
242   10.1093/molbev/msq165

243   Quinlan AR, and Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing
244   genomic features. *Bioinformatics* 26:841-842. 10.1093/bioinformatics/btq033
245   Rodriguez F, and Arkhipova IR. 2018. Transposable elements and polyploid evolution
246   in animals. *Curr Opin Genet Dev* 49:115-123. 10.1016/j.gde.2018.04.003
247   Szabo L, and Salzman J. 2016. Detecting circular RNAs: bioinformatic and experimental
248   challenges. *Nat Rev Genet* 17:679-692. 10.1038/nrg.2016.114
249   Westholm JO, Miura P, Olson S, Shenker S, Joseph B, Sanfilippo P, Celniker SE,
250   Graveley BR, and Lai EC. 2014. Genome-wide analysis of drosophila circular
251   RNAs reveals their structural and sequence properties and age-dependent neural
252   accumulation. *Cell Rep* 9:1966-1980. 10.1016/j.celrep.2014.10.062
253   Yahav T, and Privman E. 2019. A comparative analysis of methods for de novo
254   assembly of hymenopteran genomes using either haploid or diploid samples. *Sci
255   Rep* 9:6480. 10.1038/s41598-019-42795-6
256   Zhang XO, Dong R, Zhang Y, Zhang JL, Luo Z, Zhang J, Chen LL, and Yang L. 2016.
257   Diverse alternative back-splicing and alternative splicing landscape of circular
258   RNAs. *Genome Res* 26:1277-1287. 10.1101/gr.202895.115
259

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

Figure 1

An overview of the CircParser pipeline

An overview of the CircParser pipeline: Figure 1A: The pipeline includes merging of the circRNAs with overlapping genome coordinates and presents the number of different circRNAs originating from one host gene. Figure 1B: CircParser includes the prediction of circRNA structural components using a genome annotation gff/gff3 file
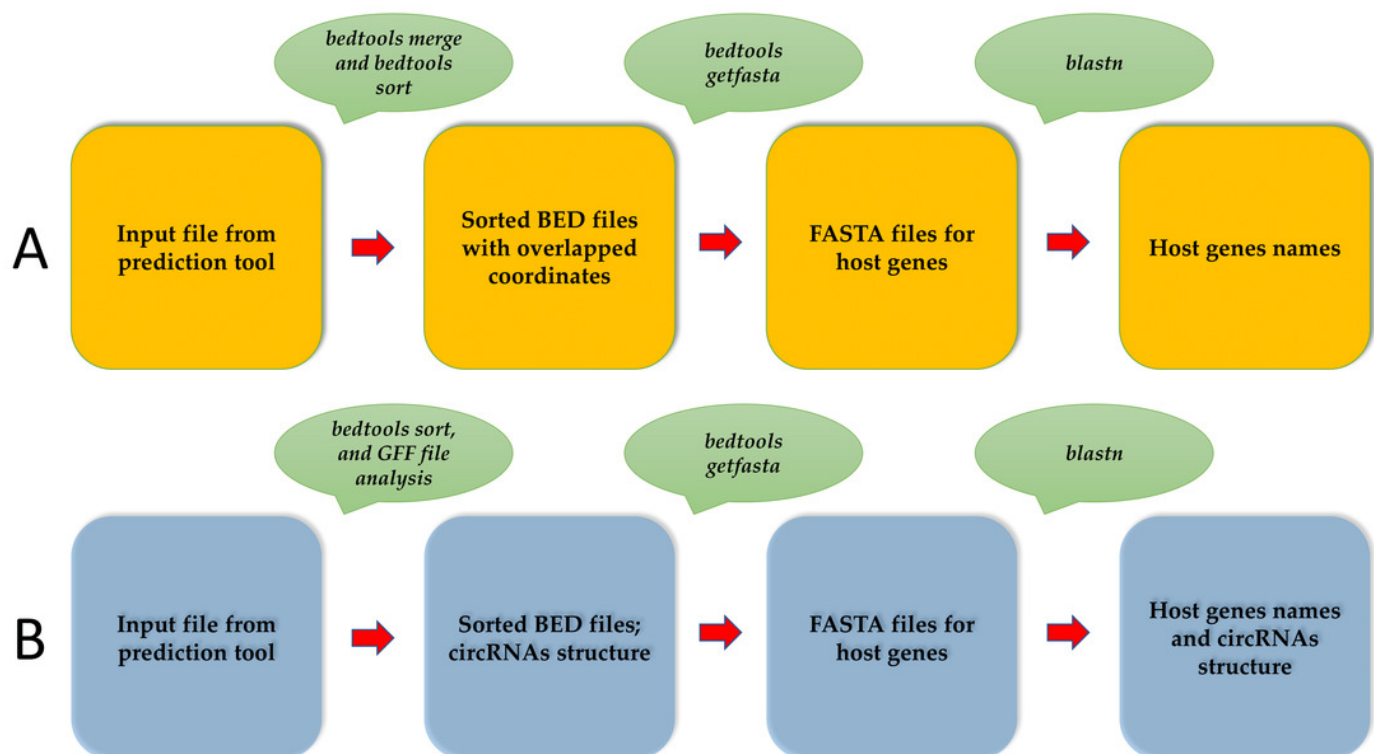
Figure 2

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

Number of circular RNAs that have been predicted by CIRI, CIRI2, CircExplorer2, find_circ, circFinder, and that are common between all prediction algorithms
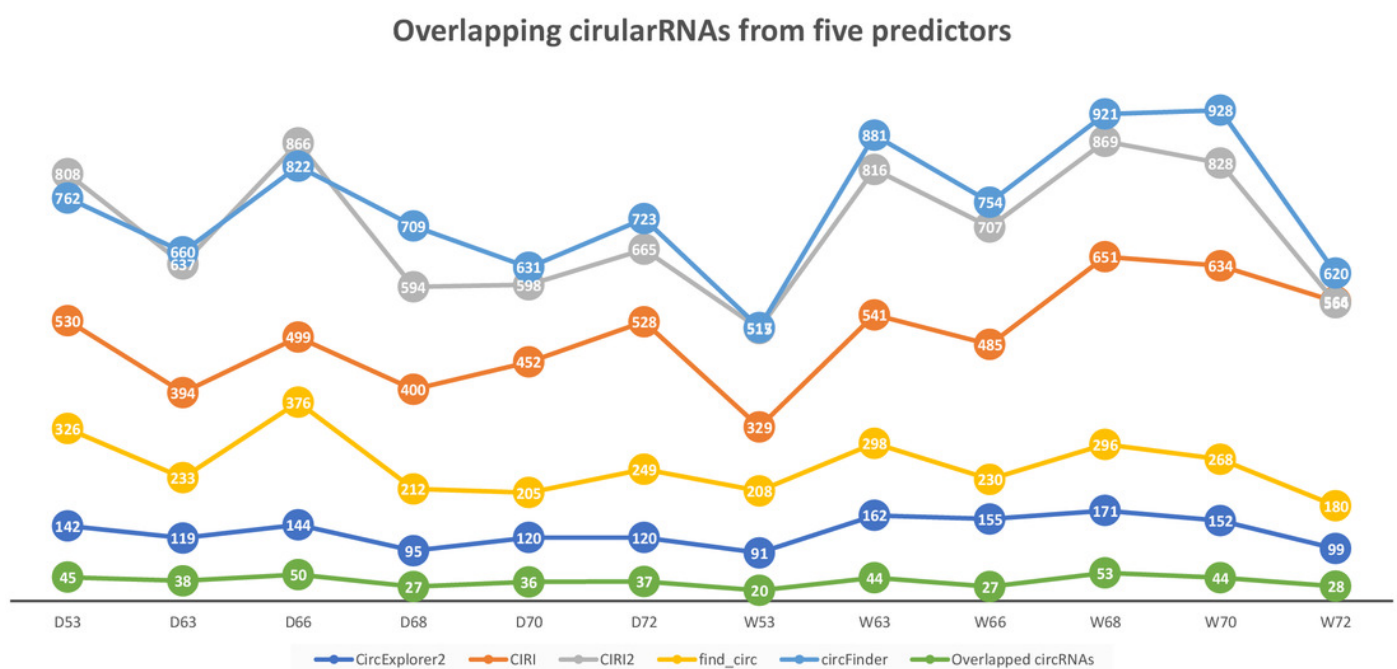
Figure 3

CircParser capacity: number of host genes that were predicted by CircExplorer2 and CircParser

The number of predicted host genes for Nile tilapia datasets

Dataset	CircParser	CircExplorer2
D53	127	77
D63	112	58
D66	120	72
D68	83	40
D70	110	73
D72	107	56
W53	83	45
W63	138	88
W66	118	72
W68	151	84
W70	141	77
W72	87	43

■ CircParser   ■ CircExplorer2