User Manual for CircPro

Last updated: 18-July-2017

Author: Xianwen Meng

From: Zhejiang University

Email: xianwen@zju.edu.cn

Description

This is a brief tutorial to describe the usage of CircPro. CircPro is an integrated pipeline designed to detect circRNAs with protein-coding potential from high-throughput sequencing data. CircPro realizes *de novo* detection of circRNAs and their protein-coding potential. In addition, it can identify circRNA isoforms and provide the sequences of circRNAs and circRNA-derived proteins. As a user-friendly tool, CircPro will facilitate the investigation of circRNA translation as well as novel functions of circRNAs and circRNA-derived proteins.

CircPro is composed of three modules. In the first module, The total/poly(A)- RNA sequencing reads are mapped to reference genome using BWA-MEM. The generated SAM alignment is used by CIRI2 for de novo detection of circRNAs. In the second module, CircPro extracts circRNA sequences. Then, CPC is used to assess their protein-coding potential. In the third module, for Ribo-Seq reads, adaptors are removed and rRNA reads are filtered. The cleaned reads are mapped to reference genome and the unmapped reads are further mapped to the library of circRNA junction sites, which is constructed by extracting N nucleotides from both sides of the junction site (N is the length of Ribo-Seq reads). Finally, CircPro generates a final circRNA list.

Installation

Several dependencies are required to run CircPro. Please make sure perl 5.10 or higher version have been installed in your computer and use Mac OS X or Linux operation system.

1. Bio::Perl See http://search.cpan.org/~cjfields/BioPerl-1.007001/Bio/Perl.pm.
2. Bio::SeqIO See http://search.cpan.org/dist/BioPerl/Bio/SeqIO.pm.
3. BWA See https://sourceforge.net/projects/bio-bwa/files/.
4. Bowtie2 See https://sourceforge.net/projects/bowtie-bio/files/bowtie2/.
5. SAMtools See http://www.htslib.org/.
6. NCBI BLAST See http://www.ncbi.nlm.nih.gov/blast/.
7. FASTX-Toolkit See http://hannonlab.cshl.edu/fastx_toolkit/index.html.

Install all the dependencies and then download and install CircPro.tar.gz. CIRI2 v2.0.5 and CPC packages are included in CircPro.tar.gz.

Run:

tar –xvzf CircPro.tar.gz

cd CircPro

sh install_cpc.sh

Currently, the protein database used by CPC was downloaded from UniProt. NCBI nr is also available for CPC. The database should be named as “prot_db”, and put under /cpc-0.9-r2/data/.

Usage

How to run CircPro:

perl CircPro.pl -c TotalRNASeq.fastq -o OutputDir -ref Genome.fa -g Gene.gtf -m ve -r RiboSeq.fastq -rRNA rRNA.fa -a Adaptor -t num -l N

The arguments of CircPro are as followings:

-C, --circ_in

FASTQ file from total/poly(A)- RNA-Seq. Paired-end FASTQ files should be separated by ",", e.g. "-C file_1.fastq,file_2.fastq". Multiple FASTQ files should be separated by ":", e.g. "-C file1.fastq:file2.fastq:file3.fastq".

-R, --ribo_in

FASTQ file from Ribo-Seq. Paired-end FASTQ files should be separated by ",", e.g. "-R file_1.fastq,file_2.fastq". Multiple FASTQ files should be separated by ":", e.g. "-R file1.fastq:file2.fastq:file3.fastq".

-O, --out

output dir

-ref, --ref_file

reference genome sequence in FASTA format

-rRNA, --rRNA_file

rRNA sequence in FASTA format (optional), it can be downloaded from Rfam or Ensembl

-G, --gene_anno

input GTF/GFF3 formatted annotation file name

-A, --adaptor

adaptor string (optional)

-H, --help

show help information

-M, --model

CNCI classification model ("ve" for vertebrate species, "pl" for plant species)

-T, --thread_num

number of threads for parallel running (default: 1)

-L, --overlap_len

minimal overlap length between Ribo-Seq reads and junction region (in each direction) (default: 5)

Example

All the packages have been tested on Ubuntu 14.04 platform, and should work on similar Linux system.

The versions of necessary packages used in this study are listed here:

- BWA (https://sourceforge.net/projects/bio-bwa/files/) version: 7.0.5a
- Bowtie2 (https://sourceforge.net/projects/bowtie-bio/files/bowtie2/) version: 2.1.0
- SAMtools (http://www.htslib.org/) version: 1.4.1
- NCBI BLAST (http://www.ncbi.nlm.nih.gov/blast/) version: 2.2.26
- FASTX-Toolkit (http://hannonlab.cshl.edu/fastx_toolkit/index.html) version: 0.0.14

Input:

The example data can be downloaded from http://bis.zju.edu.cn/CircPro/exampledata.tar.gz. The input files include genome reference file, gene annotation file, total RNA-Seq file, Ribo-Seq file and rRNA sequence file. Put these files in the CircPro directory.

Enter the directory and type as following in your terminal:

perl CircPro.pl -c TotalRNASeq_test.fastq -o ./test -ref Genome_test.fa -g Gene_test.gtf -m ve -r RiboSeq_test.fastq -rRNA rRNA_test.fa -a TGGAATTCTCGGGTGCCAAGG -t 5

Output:

1. **CircPro.out**

The identified circRNA list, including the information of genomic position, coding status, ORF length and junction reads from RNA-Seq and Ribo-Seq.

circRNA	chr	start	end	strand	junction reads (RNA-Seq)	type	parent gene	junction reads (Ribo-Seq)
chr16:30664215|30666541	chr16	30664215	30666541	+	2	exonic	ENSG00000156860	2
chr16:30766526|30766779	chr16	30766526	30766779	+	3	exonic	ENSG00000103549	0
chr16:30924220|30927901	chr16	30924220	30927901	-	2	intergenic	n/a	0
chr16:71723769|71860672	chr16	71723769	71860672	-	4	intergenic	n/a	2

2. **circIsoform.out**

It includes the information of all possible circRNA isoforms.

circRNA	classification	CPC score	ORF length
chr16:30664215|30666541_1	coding	4.28062	1481
chr16:30664215|30666541_2	coding	0.831106	712
chr16:30766526|30766779_1	noncoding	-0.89349	115
chr16:30766526|30766779_2	noncoding	-1.37772	31
chr16:30924220|30927901_1	noncoding	-0.975283	1360
chr16:71723769|71860672_1	coding	3.05625	899

3. **circRNA.fa**

All circRNA sequences in FASTA format.

4. **circProtein.fa**

The circRNA-derived protein sequences in FASTA format.