CircMarker: A Fast and Accurate Algorithm for Circular RNA Detection

Xin Li, Chong Chu, Jingwen Pei, Ion Măndoiu, and Yufeng Wu*

Computer Science & Engineering Dept., University of Connecticut, Storrs, CT, USA
{xin.li,chong.chu,jingwen.pei,ion.mandoiu,yufeng.wu}@uconn.edu

Circular RNA (or circRNA) is a type of RNA which forms a covalently closed continuous loop. It is now believed that circRNA plays important biological roles in some diseases. Within the past several years, several experimental methods, such as RNase R, have been developed to enrich circRNA while degrading linear RNA. Some useful software tools for circRNA detection have been developed as well. However, these tools may miss many circRNA. Also, existing tools are slow for large data because those tools often depend on reads mapping.

In this paper, we present a new computational approach, named CircMarker, based on k-mers rather than reads mapping for circular RNA detection as shown in figure 1. The algorithm has two parts, including reference genome proprocessing and annotations (part 1) and circular RNA detection (part 2).

In part 1, CircMarker creates a table for storing the k-mers within the reference genome that are near the exon boundaries as specified by the annotations. The k-mer table is designed to be space-efficient. We only record five types of information for each k-mer, including chromosome index, gene index, transcript index, exon index and part tag. The "part tag" specifies whether a k-mer comes from the head part or the tail part of the exon.

Part 2 is divided into five steps. (1) Sequence reads processing: examine k-mers contained in a read and search for a match in the k-mer table. (2) Filtering by hit number: short exons should be fully covered by the reads more than one time. Otherwise, the reads should be within both boundaries of the hit exons. (3) Filtering by part tags: we collect part tags from start to end, and condense the tags which belong to the same exons based on the number of hits. (4) Calling circRNA: both self-circular case (single exon) and regular-circular case(multiple exon) are considered. In the regular-circular case, we consider if the exon index increases/decreases monotonically and identify the circular joint junction at the position of the first deceasing/increasing position. (5) Refining circular RNA candidates (optional): only the candidates with support number smaller than a predefined threshold will be viewed as correct one.

We use both simulated and real data for evaluation. We compared CircMarker with three other tools, including CIRI [1], Find_circ [3], and CIRCexplorer [4] in terms of the number of called circular RNA, accuracy, consensus-based sensitivity, bias and running time. The results are shown in figure 1.

- **Simulated Data.** The simulated data is generated by the simulation script released by CIRI. The reference genome is chromosome 1 in human genome (GRCh37). The annotation file is version 18 (Ensembl 73). Two different

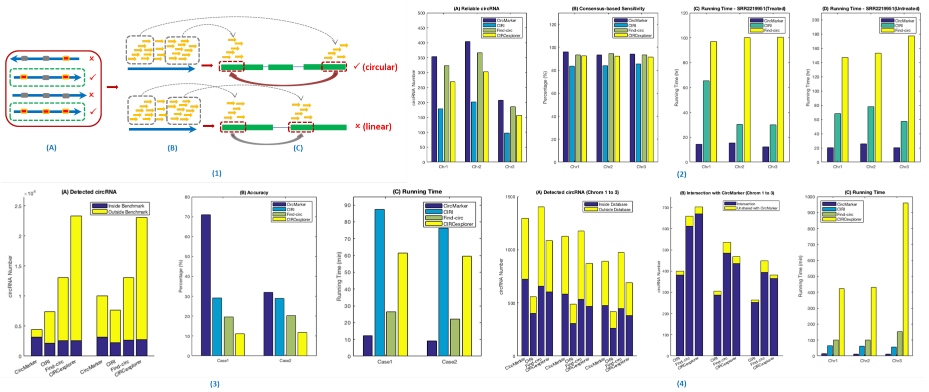

Fig. 1. High Level Approach and Results: (1) High level approach: a fast check for finding circRNA relevant reads, scanning k-mer sequentially from the beginning to the end for each read, and calling circRNA using various criteria and filters. **(2)** Results of real data based on RNase R treated/untreated reads. **(3)** Results of simulated data. **(4)** Results of real data based on RNase R treated reads with public database.

cases are simulated, including 10X circRNA & 100X linear RNA, and 50X for both circular and linear RNA.

- **Real data: RNase R treated reads with public database.** We choose CircBase [2] as the standard circRNA database of homo sapiens. The reference genome and annotation file come from homo sapiens GRCm37 version 75. The RNA-Seq reads are from SRR901967.
- **Real Data: RNase R treated/untreated Reads.** The reference genome and annotation file are from *Mus Musculus GRCm38 Release79*. RNase R treated/untreated reads are from SRR2219951 and SRR2185851 respectively.

The results show that CircMarker runs much faster and can find more circular RNA than other tools. In addition, CircMarker has higher consensus-based sensitivity and high accuracy/reliable ratio compared with others. Moreover, the circRNAs called by CircMarker often contain most circRNAs called by other tools in the real data we tested. This implies that CircMarker has low bias. CircMarker can be downloaded at: https://github.com/lxwgcool/CircMarker.

Bibliography

[1] Gao, Y., Wang, J., Zhao, F.: Ciri: an efficient and unbiased algorithm for de novo circular rna identification. Genome biology **16**(1) (2015) 4

[2] Glažar, P., Papavasileiou, P., Rajewsky, N.: circbase: a database for circular rnas. Rna **20**(11) (2014) 1666–1670

[3] Memczak, S., Jens, M., Elefsinioti, A., et al.: Circular rnas are a large class of animal rnas with regulatory potency. Nature **495**(7441) (2013) 333–338

[4] Zhang, X.O., Wang, H.B., Zhang, Y., Lu, X., Chen, L.L., Yang, L.: Complementary sequence-mediated exon circularization. Cell **159**(1) (2014) 134–147