CiLiQuant: quantification of RNA junction reads based on their circular or linear transcript origin

Annelien Morlion[1,2,*], Eva Hulstaert[1,2,3], Jasper Anckaert[1,2], Celine Everaert[1,4], Jo Vandesompele[1,2] and Pieter Mestdagh[1,2,*]

[1]Department of Biomolecular Medicine, Ghent University, 9000 Ghent, Belgium

[2]OncoRNALab, Cancer Research Institute Ghent (CRIG), 9000 Ghent, Belgium

[3]Department of Dermatology, Ghent University Hospital, 9000 Ghent, Belgium

[4]Lab for translational cancer genomics and bioinformatics, Cancer Research Institute Ghent (CRIG), 9000 Ghent, Belgium

*To whom correspondence should be addressed

Abstract

Distinguishing circular RNA (circRNA) reads from reads derived from the linear host transcript is a challenging task because of sequence overlap. We developed a computational approach, CiLiQuant, that determines the relative circular and linear abundance of transcripts and gene loci using backsplice and forward splice junction reads generated by existing mapping and circRNA discovery tools.

Availability & Implementation: CiLiQuant is implemented in Python, source code and documentation are freely available via GitHub (https://github.com/OncoRNALab/CiLiQuant).

Keywords: bioinformatics, pipeline, circRNA

Introduction

Circular RNAs (circRNAs) are a novel class of non-coding RNAs found in eukaryotic transcriptomes that result from a process called backsplicing during RNA maturation. In recent years, circRNAs are attracting considerable research attention and evidence of their involvement in normal development and disease has been reported [1–3]. Due to their stable, circular conformation, tissue-specific expression patterns and abundance in biofluids, circRNAs are emerging as potential biomarker candidates in minimally-invasive liquid biopsies [4–6]. As circRNAs share most of their sequence with their linear counterparts, it is impossible to distinguish linear from circular RNA reads that do not include the backsplice sequence itself. This hampers calculations of the relative contribution of circRNAs to aggregated gene counts and can obscure differential expression analyses. To this end, we developed a computational pipeline that determines the linear or ambiguous character of forward junctions and propose two strategies to determine the circular and linear contribution.

Statement of need

Several RNA sequencing library preparation methods can pick up both linear and circular RNA transcripts but contrary to the clear circular RNA origin of backsplice reads, the linear or circular origin of forward splice junctions is not always obvious. CiLiQuant classifies forward splice junction reads as linear only or ambiguous depending on the overlap with detected circRNA transcripts. By correcting the sum of junction reads for the number of unique junctions, the linear only and backsplice junction reads can be directly compared. The entire pipeline can be initiated with a single command and only requires Python and the Pandas package [7]. Unlike other strategies that determine circular-to-linear RNA ratios [8,9], the flexible input format of CiLiQuant allows the usage of junction count files from various combinations of mappers and circRNA quantification tools. This enables users to look at circRNA transcript fractions using their preferred mapping and circRNA detection strategy. Moreover, the dual approach allows users to look at the circular versus linear RNA abundance from different perspectives. The confidence interval provides the user with a level of uncertainty on the calculated fraction; e.g. a ratio of 2/2 has higher uncertainty than a ratio of 20/20 reads. As the output table also includes the original counts, the user can still impose a minimum count threshold and look at absolute count differences. Possible applications include, but are not limited to, comparing linear-to-circular RNA enrichment between tissues, biofluids, healthy and disease state; identifying genes or regions with abundant circRNA expression; discovering interesting circRNAs for biomarker purposes. Finally, the ambiguous

doi:10.20944/preprints202109.0183.v1

category can help to determine the relative level of mixed (linear and circular RNA) signal in aggregated gene counts. In case of sufficient sequencing depth, the linear only classification can be used as a starting point for differential expression analysis of linear RNA read counts only. This approach would be very similar to the current backsplice junction count differential expression for circRNAs. In conclusion, the CiLiQuant pipeline distinguishes linear only from potentially mixed junction reads and determines the circular and linear contribution in RNA sequencing data in a systematic and uniform way.

Implementation

The pipeline requires three input files in tab-separated format. Any combination of existing mapping and circRNA discovery tools can be used to generate the first two files: one file for forward splice junctions, e.g. from STAR [10], and one for backsplice junctions e.g. from CIRCexplorer2 [11]. The only requirements are that these junction files were generated from the same sequencing data and contain information about the coordinates of the junctions and their respective read counts in separate columns. The third input file should contain start and stop coordinates of the genes (or exons) of interest. More details and example input files can be found on GitHub.

For each gene, the pipeline classifies forward splice junctions as having a linear or ambiguous origin based on their overlap with detected backsplice junctions (figure 1A). A forward splice junction that starts and stops in between any detected backsplice junction is considered ambiguous because both linear and circular RNA can contribute to the read counts of this junction. In case the forward splice junctions do not (completely) fall within the start and stop of any detected backsplice junction, they are classified as linear. Using this information, our method determines the relative circular to linear RNA abundance at two levels, backsplice and gene level (figure 1B). At the individual backsplice level, backsplice reads are compared to the linear forward splice junction reads that are directly flanking the backsplice. Of note, sometimes this information is not available – either because there are no flanking reads or because those reads are classified as ambiguous as they could be derived from another circRNA. Therefore, an alternative calculation using the average of all linear only junction reads in the gene is provided as well. At the gene level, the average number of backsplice junction reads is compared to the average number of linear only junction reads per junction. For each circRNA fraction, an Agresti-Coull 95% confidence interval is calculated [12].

Availability of source code and requirements

Project name: CiLiQuant

Project home page: https://github.com/almorlio/CiLiQuant

Operating system: Platform independent

Programming language: Python

Other requirements: Python 3.6 or higher, Pandas 1.0.5 or higher

RRID: SCR_019319

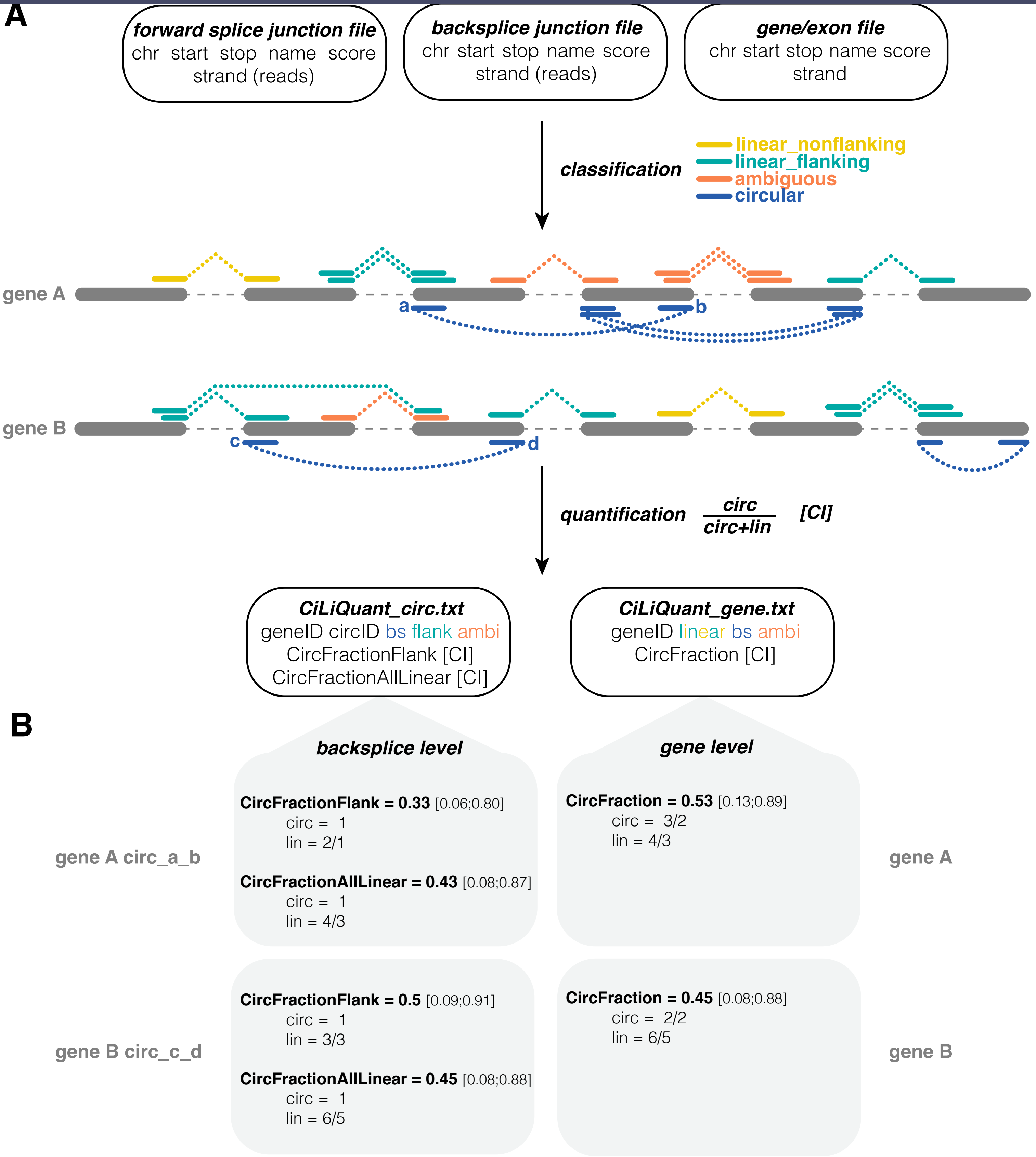
https://bio.tools/ciliquant

Acknowledgements

We want to thank Marieke Vromman for her feedback regarding different circRNA quantification tools.

Figures

Figure 1. CiLiQuant classifies splice junctions based on their linear or circular origin and determines circRNA fractions. A) Input, processing and output of the pipeline. Three input files required (number of junction reads may be in separate column or in score column), junctions are divided into four types based on overlap with detected circRNAs, quantification both at backsplice and gene level; B) Examples of circRNA fraction calculations. CircFractionFlank only considers linear junctions directly next to the backsplice of interest while CircFractionAllLinear and CircFraction consider all linear (flanking and non-flanking) junctions in the gene. In each calculation the sum of junction reads is corrected for the number of distinct junctions. Note that the counts in this example are rather low resulting in large confidence intervals. Test case with real sequencing data on GitHub. bs: backsplice; ambi: ambiguous (circular or linear origin); CI: Agresti-Coull 95% confidence interval.


A
forward splice junction file
chr start stop name score strand (reads)
backsplice junction file
chr start stop name score strand (reads)
gene/exon file
chr start stop name score strand
classification
linear_nonflanking
linear_flanking
ambiguous
circular
gene A
a
b
gene B
c
d
quantification
circ
circ+lin
[CI]
CiLiQuant_circ.txt
geneID circID bs flank ambi
CircFractionFlank [CI]
CircFractionAllLinear [CI]
CiLiQuant_gene.txt
geneID linear bs ambi
CircFraction [CI]
B
backsplice level
gene level
gene A circ_a_b
CircFractionFlank = 0.33 [0.06;0.80]
circ = 1
lin = 2/1
CircFractionAllLinear = 0.43 [0.08;0.87]
circ = 1
lin = 4/3
CircFraction = 0.53 [0.13;0.89]
circ = 3/2
lin = 4/3
gene A
gene B circ_c_d
CircFractionFlank = 0.5 [0.09;0.91]
circ = 1
lin = 3/3
CircFractionAllLinear = 0.45 [0.08;0.88]
circ = 1
lin = 6/5
CircFraction = 0.45 [0.08;0.88]
circ = 2/2
lin = 6/5
gene B

Declarations

List of abbreviations

ambi: ambiguous (circular or linear origin)

bs: backsplice

CI: Agresti-Coull 95% confidence interval

circRNA: circular RNA

Competing interests

The author(s) declare that they have no competing interests.

Funding

This work was supported by the Special Research Fund (BOF) of Ghent University [BOF19/DOC/228 to A.M., BOF19/GOA/008 to P.M., BOF16/GOA/023 to J.V.]; Fund for Scientific Research Flanders (FWO) [11C1621N to A.M., 1133120N to E.H., 1S07416N to C.E.]; Stichting Tegen Kanker (FAF-F/2018/1267); and Kom op tegen Kanker (Stand up to Cancer) the Flemish cancer society. This project was funded by the European Union's Horizon 2020 research and innovation programme under grant agreements 668858 and 826121.

Author's Contributions

CONCEPTUALIZATION: Celine Everaert, Pieter Mestdagh, Annelien Morlion, Jo Vandesompele

SUPERVISION: Pieter Mestdagh, Jo Vandesompele

PROJECT ADMINISTRATION: Annelien Morlion

INVESTIGATION: Celine Everaert, Annelien Morlion

FORMAL ANALYSIS: Annelien Morlion

SOFTWARE: Annelien Morlion

METHODOLOGY: Annelien Morlion

VALIDATION: Jasper Anckaert, Eva Hulstaert, Annelien Morlion

DATA CURATION: Annelien Morlion

RESOURCES: Jasper Anckaert

FUNDING ACQUISITION: Pieter Mestdagh, Jo Vandesompele

WRITING – ORIGINAL DRAFT PREPARATION: Eva Hulstaert, Annelien Morlion

WRITING – REVIEW & EDITING: Jasper Anckaert, Celine Everaert, Pieter Mestdagh, Jo Vandesompele

VISUALIZATION: Eva Hulstaert, Annelien Morlion

References

1. Gaffo E, Boldrin E, Molin AD, Bresolin S, Bonizzato A, Trentin L, et al.. Circular RNA differential expression in blood cell populations and exploration of circRNA deregulation in pediatric acute lymphoblastic leukemia. *Sci Rep*. Nature Publishing Group; 2019; doi: 10.1038/s41598-019-50864-z.

2. Maass PG, Glažar P, Memczak S, Dittmar G, Hollfinger I, Schreyer L, et al.. A map of human circular RNAs in clinically relevant tissues. *Journal of Molecular Medicine*. 2017; doi: 10.1007/s00109-017-1582-9.

3. Vo JN, Cieslik M, Zhang Y, Shukla S, Xiao L, Zhang Y, et al.. The Landscape of Circular RNA in Cancer. *Cell*. 2019; doi: 10.1016/j.cell.2018.12.021.

4. Salzman J, Gawad C, Wang PL, Lacayo N, Brown PO. Circular RNAs Are the Predominant Transcript Isoform from Hundreds of Human Genes in Diverse Cell Types. *PLOS ONE*. Public Library of Science; 2012; doi: 10.1371/journal.pone.0030733.

5. Su M, Xiao Y, Ma J, Tang Y, Tian B, Zhang Y, et al.. Circular RNAs in Cancer: emerging functions in hallmarks, stemness, resistance and roles as potential biomarkers. *Mol Cancer*. 2019; doi: 10.1186/s12943-019-1002-6.

6. Zhang Z, Yang T, Xiao J. Circular RNAs: Promising Biomarkers for Human Diseases. *EBioMedicine*. Elsevier; 2018; doi: 10.1016/j.ebiom.2018.07.036.

7. Jeff Reback, Wes McKinney, jbrockmendel, Joris Van den Bossche, Tom Augspurger, Phillip Cloud, et al.. pandas-dev/pandas: Pandas 1.0.5. Zenodo;

8. Ma X-K, Wang M-R, Liu C-X, Dong R, Carmichael GG, Chen L-L, et al.. CIRCexplorer3: A CLEAR Pipeline for Direct Comparison of Circular and Linear RNA Expression. *Genomics Proteomics Bioinformatics*. 2019; doi: 10.1016/j.gpb.2019.11.004.

9. Rybak-Wolf A, Stottmeister C, Glažar P, Jens M, Pino N, Giusti S, et al.. Circular RNAs in the Mammalian Brain Are Highly Abundant, Conserved, and Dynamically Expressed. *Molecular Cell*. 2015; doi: 10.1016/j.molcel.2015.03.027.

10. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al.. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*. 2013; doi: 10.1093/bioinformatics/bts635.

11. Zhang X-O, Dong R, Zhang Y, Zhang J-L, Luo Z, Zhang J, et al.. Diverse alternative back-splicing and alternative splicing landscape of circular RNAs. *Genome Res*. 2016; doi: 10.1101/gr.202895.115.

12. Brown LD, Cai TT, DasGupta A. Interval Estimation for a Binomial Proportion. *Statist Sci*. Institute of Mathematical Statistics; 2001; doi: 10.1214/ss/1009213286.