Bioinformatics, 32(7), 2016, 1094–1096
doi: 10.1093/bioinformatics/btv656
Advance Access Publication Date: 9 November 2015
Applications Note

OXFORD

Sequence analysis

Specific identification and quantification of circular RNAs from sequencing data

Jun Cheng, Franziska Metge and Christoph Dieterich*,†

Max Planck Institute for Biology of Ageing, Joseph-Stelzmann-Strasse 9B, 50931 Cologne, Germany

*To whom correspondence should be addressed.
†Present address: Section of Bioinformatics and Systems Cardiology, Department of Internal Medicine III and Klaus Tschira Institute for Computational Cardiology, University Hospital Heidelberg, Germany.
Associate Editor: John Hancock

Received on June 17, 2015; revised on November 2, 2015; accepted on November 4, 2015

Abstract

Motivation: Circular RNAs (circRNAs) are a poorly characterized class of molecules that have been identified decades ago. Emerging high-throughput sequencing methods as well as first reports on confirmed functions have sparked new interest in this RNA species. However, the computational detection and quantification tools are still limited.
Results: We developed the software tandem, `DCC` and `CircTest`. `DCC` uses output from the STAR read mapper to systematically detect back-splice junctions in next-generation sequencing data. `DCC` applies a series of filters and integrates data across replicate sets to arrive at a precise list of circRNA candidates. We assessed the detection performance of `DCC` on a newly generated mouse brain data set and publicly available sequencing data. Our software achieves a much higher precision than state-of-the-art competitors at similar sensitivity levels. Moreover, `DCC` estimates circRNA versus host gene expression from counting junction and non-junction reads. These read counts are finally used to test for host gene-independence of circRNA expression across different experimental conditions by our R package `CircTest`. We demonstrate the benefits of this approach on previously reported age-dependent circRNAs in the fruit fly.
Availability and implementation: The source code of `DCC` and `CircTest` is licensed under the GNU General Public Licence (GPL) version 3 and available from https://github.com/dieterich-lab/ [DCC or CircTest].
Contact: christoph.dieterich@age.mpg.de
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

CircRNAs are a recently rediscovered class of abundant and universally expressed RNA species (Jeck and Sharpless, 2014). The development of software solutions for detecting and quantifying circRNAs is slowly gaining momentum despite the increasing interest in circRNAs. CircRNAs are characterized by a back-splicing event where the downstream 3′ splicing 'tail' joins back with the upstream 5′ splicing 'head' to form a circular RNA structure (see Supplementary Fig. S1). This structure becomes visible in sequencing data as 'scrambled' exons as the co-linearity of genome and transcript is violated. However, the existence of 'back-splicing' junction reads does not prove the circularity of the transcript of origin. Other mechanisms (e.g. trans-splicing) may yield such reads, but only circRNAs are depleted in polyA-enriched samples and resistant to RNAse R treatment. So far, all available algorithms are using these 'back-splicing' reads as a key element in circRNA detection. As a consequence, detection performance heavily depends on the employed read mapper and its ability to map circRNA junction reads to the underlying genome or transcriptome. The two most recent addition to the circRNA detection tool set are `CIRI` (Gao *et al.*, 2015) and `KNIFE` (Szabo *et al.*, 2015). Other circRNA detecting pipelines have been published along with research papers but seem to perform worse than the two in the corresponding benchmarks.

It has been recently shown that circRNAs may vary in their abundance relative to their host gene (Westholm *et al.*, 2014). It is interesting to statistically discern cases where circRNA and linear isoform expression are tightly coupled from the ones where expression diverges between the two. To this end, read counts over the back-splice junction are typically used as an estimate of circRNA expression and read counts from exons or linear splice junctions as an estimate of the host gene expression.

2 Results

2.1 Efficient circular RNA detection by DCC

We motivate our circRNA detection workflow from the review paper of Lasda and Parker (2014). Essentially, we are looking for chimeric read alignments, which are reported by the fast and splice-aware read mapper STAR (Dobin *et al.*, 2013). Our method works with single and paired-end data and executes a sequence of analysis steps as outlined in Supplementary Figure S2 and below:

1. If paired end reads are available, mapping of mates must be consistent with a circular RNA template.
2. Filtering by minimal number of junction reads per biological replicates to suppress 'noisy' candidates.
3. Test for presence of canonical GT/AG splicing signal at circRNA junction borders. Candidates with non-canonical splicing signals are discarded.
4. Removal of back-splicing events that map to the mitochondrial genome.
5. Suppression of mapping artefacts by masking candidates from repetitive or homologous regions.
6. We report new alternative linear splicing events (circle skipping junctions) if detected (see Supplementary Fig. S8).

As a result set, we report identified candidate circles by genomic coordinates, gene locus annotation, type of circularization and overlap with gene body annotation (see Supplementary Tables S1–S3).

2.2 Beta-Binomial for modelling circular RNA expression

We employ a beta-binomial model downstream of the circRNA detection and read counting to model changes in circRNA expression relative to that of the host gene. We assume that the observed read counts either originate from linear or circular isoforms. Suppose we have k groups (e.g. age groups) and each group has m replicates. We define $\{(x_{ij}, n_{ij}) : i = 1, \ldots, m; j = 1, \ldots, k\}$ as the set of relevant read counts for a given candidate circular RNA junction where x_{ij} is the number of back-spliced reads and n_{ij} is the number of all reads (circRNA plus linear plus pre-mRNA). We assume that x_j (the number of circular junction read counts for group j) follows a binomial distribution with n trials and probability of success θ. We further assume that the parameter θ of the binomial distribution follows a beta distribution with pseudocounts α and β. This leads to an analytically tractable compound distribution where one can think of the parameter of the binomial distribution as being randomly drawn from a beta distribution (Skellam, 1948):

$$\theta \sim \text{Beta}(\alpha, \beta); x_j \sim \text{Bin}(n, \theta)$$

We are interested in testing whether the mean $\mu = \alpha(\alpha + \beta)^{-1}$ of the distribution differs between the respective groups $j \in 1, \ldots, k$. Our null hypothesis states that there is no difference between group means $\mu_1 = \mu_2 = \ldots = \mu_k$ while overdispersion is the same for all groups. The alternative hypothesis states that there is heterogeneity across group means. We select between the two alternative hypotheses with a likelihood ratio test (ANOVA). All relevant routines are implemented in the R package `CircTest`, which depends on the `aod` packages Lesnoff *et al.* (2012).

2.3 Benchmark of DCC

We contrast the performance of DCC with CIRI and KNIFE on several RNA-seq datasets (see Data section in Supplementary Text). In principle, true positive circRNAs are resistant to RNase R treatment, and thus the same circRNA should be detectable from RNase R treated samples as well as from rRNA-depleted total RNA-seq samples. Especially, the CircleSeq protocol (Jeck and Sharpless, 2014) provides a gold standard as it involves RNase R treatment, which will partially digest linear isoforms, while circRNAs are preferentially retained. For simplicity, we define circRNA candidates that were predicted from both experiments (CircleSeq and rRNA-depleted) as true positives (TP), whereas candidates that are just predicted from rRNA-depleted RNA-seq libraries were defined as false positives (FP). We acknowledge the fact that incomplete RNAse R digestion of linear isoforms may lead to mis-classifications in this simple scheme. Figure 1A shows the benchmark results for RNA-seq data from whole mouse brain. DCC predicts fewer circRNA candidates than KNIFE. However, the precision (TP/(TP + FP)) of DCC is much higher (97%) than the one of CIRI and KNIFE with 89% and 41%, respectively. Intriguingly, the total number of true positive predictions from DCC (2744) is highest on this particular data set. Similar results were obtained for another dataset (see Supplementary Text and Supplementary Fig. S6). Supplementary Figure S4 provides more details on prediction performance and overlap. DCC is implemented in python 2.7, was extensively tested on Mac OS and Linux systems and runs considerably faster than CIRI (see Supplementary Fig. S3).

2.4 Host gene independent regulation of circRNA abundance with development

We used DCC subsequently to study circRNA abundance in animal development and quantified relative changes of circRNA versus host gene expression based on read count data (see Supplementary Fig. S1) with our `CircTest` package. Briefly, we used the rRNA-depleted total RNA-seq data from Westholm *et al.* (2014) and focused on brain samples from 3 developmental stages (1 day, 4 days and 20 days) across 6 biological replicates. We tested circRNA candidates that were present in at least 6 out of 18 samples with at least 5 junction spanning reads for divergent expression to the host gene. From the initial 116 circRNA candidates, 72 of the circRNAs were significant (P-value ≤ 0.05 after correction for multiple testing, see Supplementary Table S4).

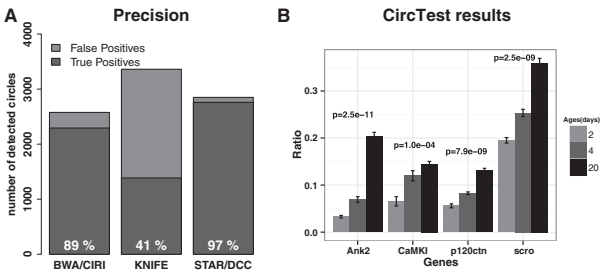

Fig. 1. (**A**) Performance of DCC compared with CIRI and KNIFE on paired-end sequencing data from whole mouse brain. (**B**) Differentially circRNA expression for four selected candidates. Ratio of circular junction read counts to average total counts at exon borders are shown along with the P-values from `CircTest`

Westholm *et al.* (2014) reported on relative circRNA expression changes as measured by qPCR ratio for four selected genes. Three of them (*Ank2*, *scro*, *p120ctn*) ranked top in our list of significant host-independent circRNAs and *CaMKI* ranked 22th (Fig. 1B and Supplementary Table S4). Our results show the same age-dependant increase as in Westholm *et al.* (2014), Figure 7E.

3 Discussion

We present a software tandem for circRNA detection (`DCC`) and relative quantification (`CircTest`). We could show that `DCC` has a higher precision than its competitors for single and paired-end data sets. Moreover, we adapt a statistical framework based on the beta-binomial distribution for identifying host gene independent changes in circRNA expression. Additional information are provided in Supplementary Text.

Acknowledgement

Acknowledgements are part of the Supplementary text.

Funding

We thank the Max Planck Society for supporting our research.

Conflict of Interest: none declared.

References

Dobin,A. *et al.* (2013) STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, **29**, 15–21.

Gao,Y. *et al.* (2015) CIRI: an efficient and unbiased algorithm for de novo circular RNA identification. *Genome Biol.*, **16**, 4.

Jeck,W.R. and Sharpless,N.E. (2014) Detecting and characterizing circular RNAs. *Nat. Biotechnol.*, **32**, 453–461.

Lasda,E. and Parker,R. (2014) Circular RNAs: diversity of form and function. *RNA*, **20**, 1829–1842.

Lesnoff,M. and Lancelot,R. (2012) *aod: Analysis of Overdispersed Data*. R package version 1.3.

Skellam,J.G. (1948) A probability distribution derived from the binomial distribution by regarding the probability of success as variable between the sets of trials. *J. R. Stat. Soc. Ser. B (Methodological)*, **10**, 257–261.

Szabo,L. *et al.* (2015) Statistically based splicing detection reveals neural enrichment and tissue-specific induction of circular RNA during human fetal development. *Genome Biol.*, **16**, 126.

Westholm,J.O. *et al.* (2014) Genome-wide analysis of Drosophila circular RNAs reveals their structural and sequence properties and age-dependent neural accumulation. *Cell Rep.*, **9**, 1966–1980.