Sensitive, reliable, and robust circRNA detection from RNA-seq with CirComPara2

Enrico Gaffo[1,*], Alessia Buratin[1,2], Anna Dal Molin[1], Stefania Bortoluzzi[1,3,*]

[1] Department of Molecular Medicine, University of Padova, Padova, Italy

[2] Department of Biology, University of Padova, Padova, Italy

[3] Interdepartmental Center For Innovative Biotechnologies, University of Padova, Padova, Italy

[*] corresponding authors

Tel: +39-049-827-6502

Fax: +39-049-827-6208

Email: stefania.bortoluzzi@unipd.it (S.B.), enrico.gaffo@unipd.it (E.G.)

Running title: CircRNA detection from RNA-seq with CirComPara2

Keywords: circRNAs, bioinformatics, transcriptomics, RNA-seq

Abstract

Circular RNAs (circRNAs) are a large class of covalently closed RNA molecules that originate by a process called back-splicing. CircRNAs are emerging as functional RNAs involved in the regulation of biological processes as well as in disease and cancer mechanisms. Current computational methods for circRNA identification from RNA-seq experiments are characterised by low discovery rates and performance dependent on the analysed data set. We developed a new automated computational pipeline, CirComPara2 (https://github.com/egaffo/CirComPara2), that consistently achieves high recall rates without losing precision by combining multiple circRNA detection methods. In our benchmark analysis, CirComPara2 outperformed state-of-the-art circRNA discovery tools and proved to be a reliable and robust method for comprehensive transcriptomics characterisation.

Introduction

Recent research uncovered that eukaryotic transcriptomes comprise thousands of stable circular RNAs (circRNAs) that originate by a process called back-splicing, where the transcript 3' and 5' ends are covalently joined (Xiao et al. 2020). Rather than being transcriptional byproducts, circRNA molecules exert critical functions in cell biology through different mechanisms (Bonizzato et al. 2016). By interacting with microRNAs, circRNAs can regulate gene expression and govern important oncogenic axes (Hansen et al. 2013); moreover, similarly to long non-coding RNAs, they can control diverse cellular processes by decoying RNA-binding proteins and scaffolding molecular complexes (Du et al. 2016). CircRNAs can also function as templates for translation to encode functional peptides (Xiao et al. 2020)(Wu et al. 2021b) and regulate the transcription of their parental gene (Li et al. 2015). Nowadays, circRNAs are considered key players that can impact all the cancer hallmarks (Hanniford et al. 2020; Slack and Chinnaiyan 2019). The discovery of circRNA regulatory roles and their potential as biomarkers given by higher stability compared to linear RNAs (Rajappa et al. 2020) have actuated the integration of circRNA investigation in conventional transcriptomics, especially in cancer research and studies of pathological conditions (Santer et al. 2019; Hua et al. 2019), including

viral infections (Awan et al. 2021).

Studies of circRNAs rapidly increased in pace thanks to the development of bioinformatics tools that identify the sequences spanning circRNA back-splice junctions from total RNA-seq data. To date, several methods for circRNA identification have been developed (Chen et al. 2020). Still, none outperforms the others since they all provide either highly sensitive or highly precise predictions and highly variable performance across benchmark data sets (Chen et al. 2015; Hansen et al. 2016; Zeng et al. 2017). Interestingly, Hansen (Hansen 2018) observed that circRNA detection methods largely agreed on true predictions, whereas circRNAs identified by single methods were enriched in false-positive guesses (FPs) and suggested selecting circRNAs commonly predicted by two or more methods to obtain dependable results.

We formerly implemented CirComPara (Gaffo et al. 2017), an automated computational pipeline combining four circRNA detection methods, including CIRCexplorer (Zhang et al. 2014), CIRI2 (Gao et al. 2018), Findcirc (Memczak et al. 2013) and Segemehl (Hoffmann et al. 2014). CirComPara controlled the FP number by considering only the circRNAs commonly detected by two or more methods.

In CirComPara2, we have considerably improved our tool by: (i) including five additional circRNA detection methods, (ii) updating the software of the already integrated tools, (iii) implementing a more accurate counting of the back-spliced reads, (iv) increasing the analysis pipeline flexibility, and (v) including the procedure to calculate the linear expression related to the circRNAs.

In this work, we first show that nine widely used circRNA detection tools could miss circRNAs of interest. Then, we confirm that CirComPara2 correctly reports circRNAs overlooked by other methods and achieves significantly higher sensitivity with no loss of precision. Finally, assessment on simulated data and 142 public RNA-seq samples demonstrated the consistent higher performance of CirComPara2 compared with state-of-the-art methods.

Results

CircRNA detection methods could miss abundant circRNAs.

We simulated RNA-seq expression data of 5,680 circRNAs from the whole human genome ("simulated data set"; see Methods) to evaluate the characteristics of circRNA detection methods' false-negative predictions (FNs), i.e., true circRNAs not identified as such. We applied nine widely used computational pipelines for circRNA discovery including circRNA_finder (Westholm et al. 2014), CIRI2 (Gao et al. 2018), DCC (Cheng et al. 2016), Findcirc (Memczak et al. 2013), Segemehl (Hoffmann et al. 2014), and CIRCexplorer2 (Zhang et al. 2016). CIRCexplorer2 was coupled to any of BWA (Li 2013) (C2BW), Segemehl (C2SE), STAR (Dobin et al. 2013) (C2ST) and TopHat-Fusion (Kim and Salzberg 2011) (C2TH) aligners, thus composing four different pipelines.

On average, 49% of FNs detected by each method showed higher expression than the overall circRNA median expression (Figure 1a), suggesting that nearly half of the missed circRNAs had a considerable expression level no matter which method was applied. Besides, the expression distribution of the FNs was similar to the correctly identified circRNAs, i.e., the true-positive findings (TPs), whereas the false-positive (FP) expression was generally low.

4

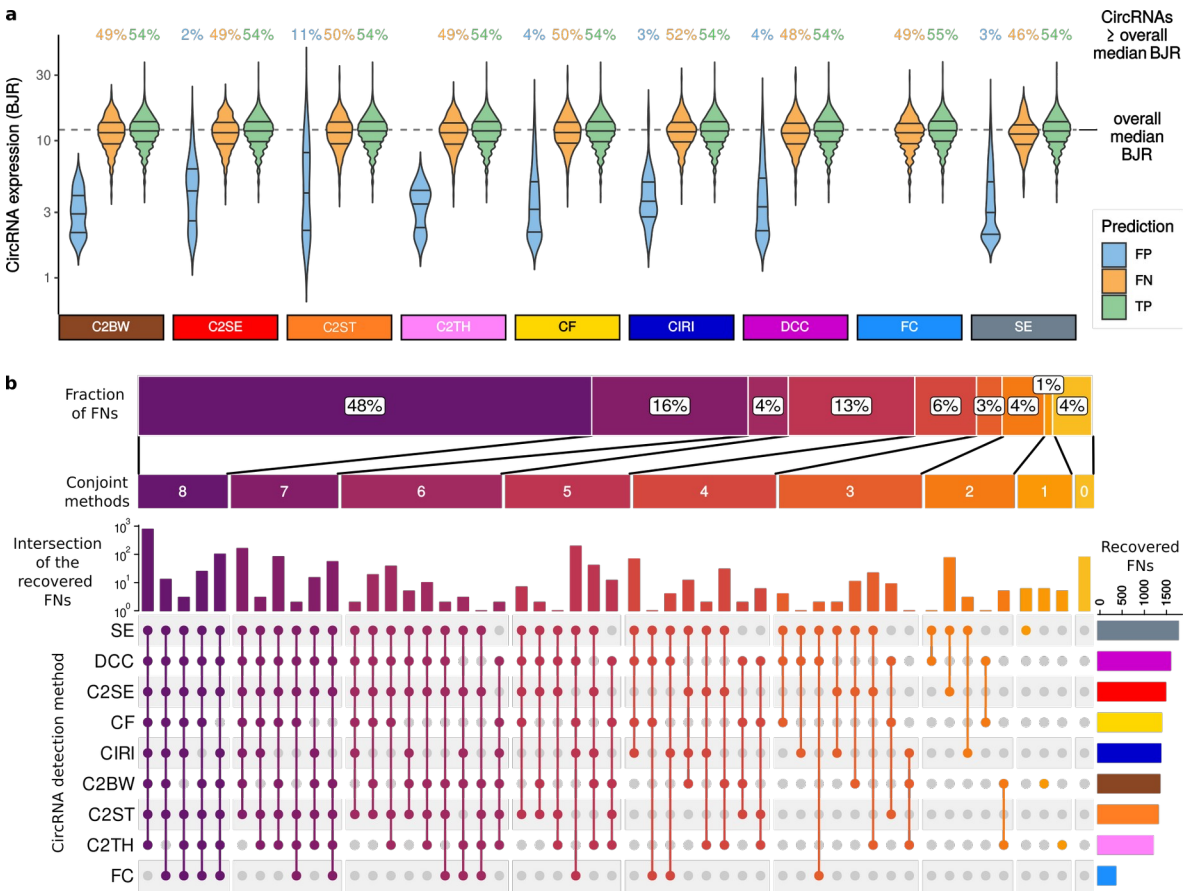

Figure 1. CircRNAs overlooked by common detection methods and the CirComPara2 approach. a, The expression level of predicted circRNAs. BJR: back-splice junction read counts; C2BW: CIRCexplorer2 on BWA; C2SE: CIRCexplorer2 on Segemehl; C2ST: CIRCexplorer2 on STAR; C2TH: CIRCexplorer2 on TopHat-Fusion; CF: circRNA_finder; CIRI: CIRI2; FC: Findcirc; SE: Segemehl; FP: false-positive; FN: false negative; TP: true positive. **b**, Number of methods detecting circRNAs not detected by other methods and the number of the FNs detected. Colour refers to the number of methods conjointly detecting circRNAs missed by other tools. The vertical bars show the number of FN circRNAs detected by the methods indicated in the coloured dots below the bars. The bars denote disjoint circRNA sets. Grey dots indicate the methods failing to detect the circRNAs considered in the bar chart on the top. The horizontal bars on the right represent the overall number of FN circRNAs detected by the methods. The horizontal bar shows the percentage of detected FNs by grouping method combinations according to the number of combined methods.

Multiple methods compensate each other, and method combinations increase the detection sensitivity.

We further examined the 1,945 circRNAs undetected by one or more tools, referred to as the "FN set" from now on, by counting how many circRNAs in the FN set each method could detect.

Interestingly, only 4% circRNAs of the FN set were undetected by all methods (Figure 1b), whereas

5

96% could be identified at least by one among the nine tools. Specifically, 1% FNs were detected individually by Segemehl, C2BW, and C2TH; and 95% were commonly identified by various combinations of two or more tools. Almost half the FNs (48%) were conjointly detected by eight out of nine methods, with Segemehl, DCC and C2SE providing the most inclusive predictions. Instead, Findcirc showed the least number of recovered FNs. However, no method entirely covered the predictions of Findcirc, indicating some specificity of its algorithm.

Overall, this analysis suggested that algorithms with possibly different and complementary features can compensate each other and improve the detection rate if applied together.

Workflow and features of CirComPara2

Following the observation reported in the previous paragraph, we enhanced CirComPara by including more circRNA detection methods to improve its sensitivity. Moreover, we introduced new features that made CirComPara2 more flexible, computationally efficient, and resilient.

CirComPara2 implements a fully automated computational pipeline for circRNA detection, quantification, annotation, and integration with linear gene expression data (Figure 2a). Several parameters are available to customise the analysis workflow and the integrated methods. The minimal input consists of the RNA-seq reads in FASTQ format and a reference genome in FASTA format. Optionally, the user can also provide the gene annotation in GTF format. The software will then build the genome indexes for each read aligner and perform the necessary file format conversions. Previously computed indexes can be reused as input to save computing time.

6

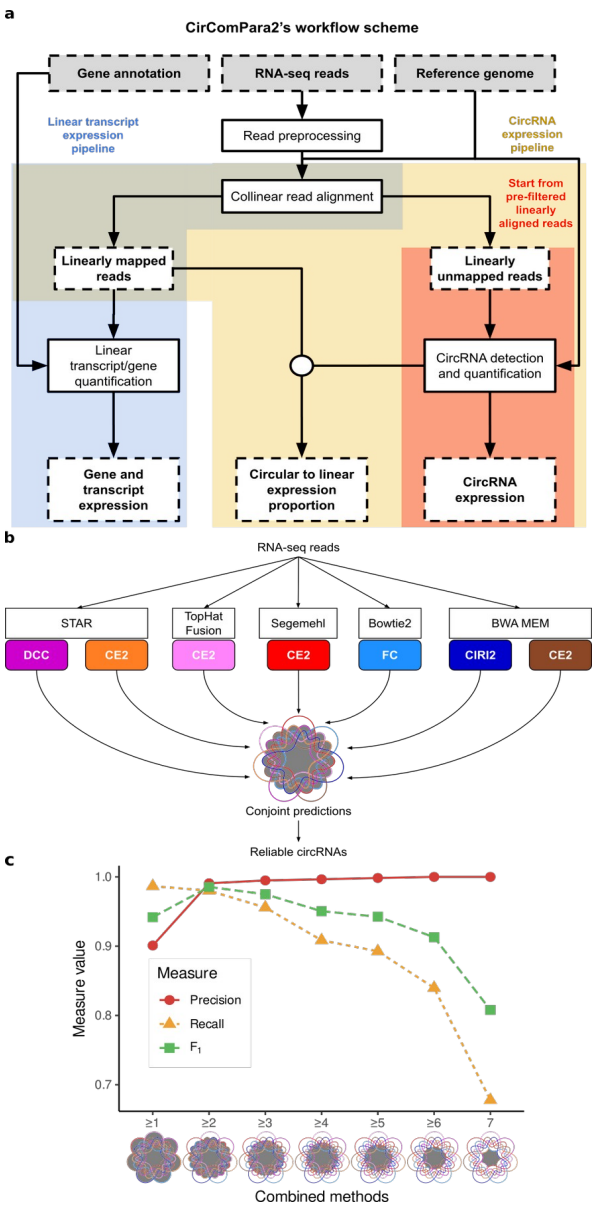

Figure 2. The CirComPara2 workflow and approach. a, Boxes with dashed contour indicate input and output data; boxes with solid-line contour indicate computing tasks; the circular connector indicates merging data. Background colours highlight the different pipeline branches (linear transcript analysis in blue, full circRNA analysis in yellow, and strict circRNA analysis in red). **b**, Detail of the CirComPara2 strategy ("CircRNA detection and quantification" box in **a**) showing the embedded circRNA detection methods (coloured rounded corner boxes) with the respective chimeric read aligners (white boxes). The central Venn diagram represents the optimised combination of method prediction intersections implemented by CirComPara2 with circRNAs conjointly detected by two or more methods (grey filled intersections) retained as default. Method abbreviations as in Figure 1. **c**, Precision (red), recall (yellow) and F_1-score (green) for different numbers of methods considered by CirComPara2: ≥1 indicates the union of all single methods predictions, ≥2 indicates any pair of methods, and so on, up to 7, which indicates the conjoint predictions of all the methods considered. The grey filled parts in Venn diagrams show the intersections considered by the conjoint method combinations.

The CirComPara2 workflow comprises an optional pre-processing of the input raw reads by Trimmomatic (Bolger et al. 2014) to trim or discard low-quality reads. Statistics of the preprocessing steps are produced with the FastQC tool (Andrews 2010). Next, the reads are aligned collinearly to the reference genome using HISAT2 (Kim et al. 2019) to (i) identify the reads that are later used for linear transcript analysis (Figure 2a) and (ii) extract the reads not collinearly aligned, which are used to detect back-splices. The linear gene and transcript expression analysis is performed with StringTie (Pertea et al. 2015) and produces files that can be easily imported into packages for downstream expression analysis such as *tximport* (Soneson et al. 2015) and *tximeta* (Love et al. 2020). The

7

circRNA analysis aligns the collinearly unmapped reads independently with five methods allowing chimeric alignments, namely Bowtie2 (Langmead and Salzberg 2012), BWA-MEM, Segemehl, STAR and TopHat-Fusion (Figure 2b). The chimeric aligner outputs are subsequently parsed by six circRNA detection tools, which compose the nine different circRNA detection sub-pipelines presented in the previous paragraph. Of note, the computationally expensive chimeric alignment step is performed only once per aligner and reused by multiple circRNA detection tools, boosting CirComPara2 efficiency. For instance, the same alignments from STAR are passed to CIRCexplorer2, circRNA_finder, and DCC (Figure 2b). The outputs of the various tools are automatically handled, converting them into a standard format to compare the predicted back-splices. Moreover, the identifiers of the back-spliced reads are collected while keeping track of the predicting method to obtain non-redundant read counts for each circRNA. Finally, the linear expression of circRNA host genes is evaluated by counting the reads collinearly mapped at each back-splice junction using bedtools (Quinlan and Hall 2010), GNU parallel (Tange 2020), and custom scripts.

CirComPara2 has a modular and highly parallelised implementation that makes it computationally efficient and resilient. By using custom parameters, CirComPara2 allows skipping computation tasks that are not of interest to the user. For instance, the user can select to run only the pipeline branch computing the linear or the circular transcript analysis, the collinear alignments (for instance, if they were previously computed), or both the collinear alignment and the linear transcript pipeline branch, therefore performing only the circRNA detection from pre-filtered reads. Plus, the Scons (www.scons.org) engine is leveraged to run independent tasks in parallel, resume an interrupted analysis by performing only uncompleted tasks, and, if the user changed some parameters from a previous run, compute only the tasks dependent on the modified parameters.

CirComPara2 is available as standalone software (https://github.com/egaffo/circompara2) and Docker image (https://hub.docker.com/r/egaffo/circompara2), which facilitates installation, portability, and reproducibility of the analysis.

Optimal method combination for circRNA detection

Concurrently to the improvement of CirComPara2 sensitivity, we wanted to control the number of the

introduced FPs to preserve a high precision. As observed in a previous study (Hansen 2018), the combination of specific methods does not ameliorate the discovery of true circRNAs. Accordingly, our method combination assessment on the simulated data showed that combining circRNA_finder or Segemehl with other methods contributed to increasing the FP number (Supplementary Figure S1). For this reason, circRNA_finder and Segemehl were excluded in the CirComPara2 default method combination (Figure 2b); nevertheless, these two methods can be included if enabled by the user.

Next, we fine-tuned the number of conjoint methods considering the amount of recovered FNs against the introduced FPs to optimise the method combination strategy. As expected, considering the predictions from all methods resulted in the highest recall (0.99) and the lowest precision (0.90) among the combination strategies (Figure 2c). Further, excluding predictions from single methods, i.e., selecting only the circRNAs commonly detected by two or more tools, showed a slightly reduced recall (0.98) with a substantially increased precision (0.99). This large precision gain indicated that most of the FPs were predicted by single methods. Further increasing the number of conjoint methods (i.e., from three-or-more to all-seven methods conjoint) lead to a considerable decrease of recall (0.96 to 0.68) with only a modest increase of precision (0.99 to 1.00).

To evenly weigh recall and precision in ranking the method combination strategies, we calculated the F_1-score for each combination. The best trade-off between recall and precision, indicated by the highest F_1-score (0.99), was obtained with the two-or-more method strategy (Figure 2c).

CirComPara2 outperforms other methods in simulated data.

We next set CirComPara2 with the two-or-more method strategy and compared it with the single methods on the simulated data. CirComPara2 obtained the highest F_1-score (0.99) by achieving the highest recall (0.98) while holding a precision comparable to or higher than the other algorithms (0.99 versus 0.92–1.00; Figure 3a), confirming that CirComPara2 rectified the true circRNAs missed by the other methods.

9

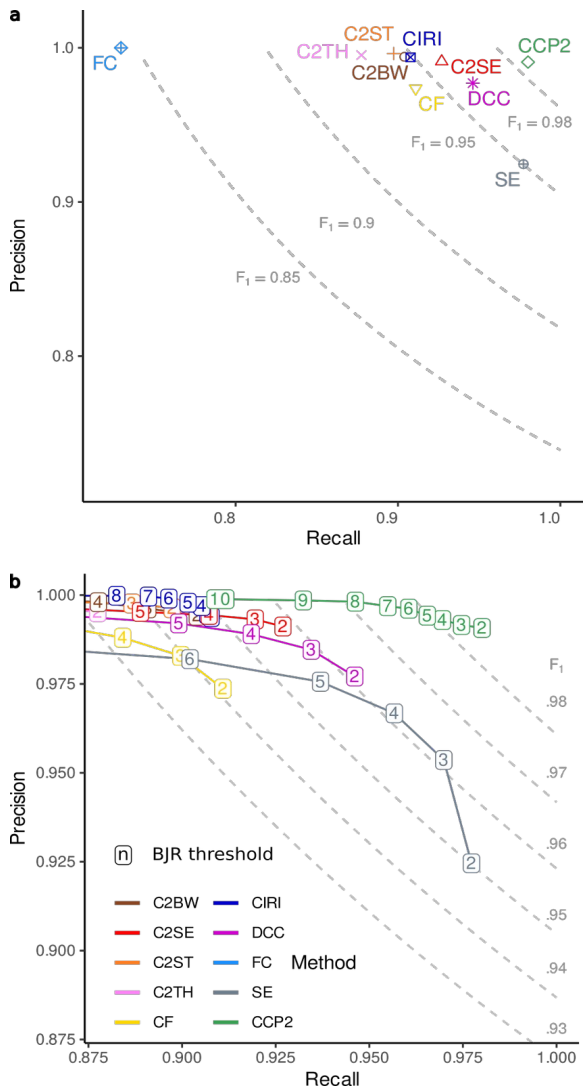

Figure 3. Performance of circRNA detection methods on the simulated data set. a, Precision and recall of the circRNA detection methods (labels as in Figure 1), including CirComPara2 (CCP2; green dot); the dashed-line curves display plot areas for 0.85, 0.90, 0.95 and 0.98 F_1-scores. **b**, CircRNA detection methods' performance upon application of filters from 2 to 10 minimum circRNA raw expression estimates (BJR: back-splice junction read counts); dashed-line curves delimit F_1-scores plot areas from 0.92 to 0.98 with 0.01 increase steps; method colours as in (a).

To assess the extent of the annotation-guided method contribution to CirComPara2 predictions, we performed the analysis also with a pruned gene annotation input to the algorithms (see Methods). As expected, the CIRCexplorer2 pipelines showed a dramatic reduction (up to -0.14) of the recall and F_1-scores (Supplementary Figure S2). Instead, CirComPara2 maintained the highest F_1-score (0.98), suggesting that it can be efficient when applied to RNA-seq data of organisms with incomplete or poor genome annotation by leveraging the embedded annotation-independent tools.

A typical circRNA expression analysis usually involves post-detection data cleaning to remove background noise signals or mapping errors and poorly expressed circRNAs of little interest (Chen et al. 2016). Consequently, the circRNAs with small back-splice junction read counts (BJRs) are routinely filtered out. As shown in Figure 1a, FPs generally have small BJRs, suggesting that more reliable circRNAs can be retained by simply filtering according to expression abundance. We applied

10

170 this procedure to our data, progressively filtering out circRNAs with less than two up to 10 BJRs. We
171 stress that removing circRNAs with ≤10 BJRs was an extreme filter since it was close to the overall
172 median BJR (Figure 1a). As expected, by increasing the minimum-BJR filter threshold, all methods
173 scored higher precision but with a corresponding reduced recall, indicating that true circRNAs were
174 discarded as well (Supplementary Figure S3). Nevertheless, CirComPara2 maintained the highest
175 recall compared to other methods at equal precision (Figure 2b; Supplementary Table S1), suggesting
176 that the circRNAs recovered by CirComPara2 were valid findings with a considerable abundance that
177 may represent relevant circRNAs in actual experiments.

178 *Benchmark real data sets*

179 The gold standard for evaluating circRNA detection methods on real RNA-seq data is to compare
180 ribosomal RNA-depleted (ribo$^-$) with circRNA-enriched sequencing libraries. The most used
181 technique is the additional treatment of the ribo$^-$ library with RNase R to exploit the exonuclease
182 degradation resistance of circRNAs derived by their lack of a single-stranded 3′ end (Hansen et al.
183 2016; Hansen 2018; Gao et al. 2018; Zeng et al. 2017; Szabo and Salzman 2016). Accordingly, we
184 collected a total of 142 public real RNA-seq data samples for which sample-matched ribo$^-$ and RNase
185 R-treated libraries were available (Table 1). This validation data set comprised samples of human cell
186 lines and various human, Rhesus macaque, and mice tissues from six independent studies.
187 The matched RNase R-treated samples were used as a control to define the TP circRNAs (see
188 Methods). Assuming that the RNase R treatment would deplete the linear more than the circular
189 transcripts, we considered TPs the circRNAs having circular-to-linear expression proportion (CLP)
190 higher or equal in the RNase R-treated compared to the ribo$^-$ matched samples according to at least
191 one method. In this way, we accounted for different sequencing library depths between the matched
192 samples. Moreover, each circRNA host-gene linear expression was estimated commonly to all the
193 detection tools and independently of the circRNA abundance estimated by each method, allowing to
194 remove possible advantages given by specific method quantification approaches.

195 **Table 1**. Real data sets of rRNA- and rRNA-/RNase R treated matched pairs of samples used to benchmark the circRNA
196 detection methods. Tissues and cell lines are from humans unless specified.

SRA/GEO/BIGD ID	Tissue(s)/cell line(s)	Study reference
SRR3476956, SRR1636985, SRR1636986; SRR3476958, SRR1637089, SRR1637090. SRR3479244, SRR3479243	HeLa, HEK293	(Gao et al. 2018) (Gao et al. 2016)
GSE130905 (10 samples)	HeLa	(Xiao and Wilusz 2019)*
CRA001838 (8 samples)	HeLa	(Zhang et al. 2020)
SRR444655, SRR444975; SRR445016, SRR444974	Hs68	(Jeck et al. 2013)
GSE113120 (24 samples)	22Rv1, 42D, PC3, V16, LNCaP	(Chen et al. 2019)
PRJCA000751 (88 samples)	*Homo sapiens* (17 tissues)	(Ji et al. 2019)
	Mus musculus (14 tissues)	
	Macaca mulatta (13 tissues)	

197 * Data set GSE130905 used an improved protocol for circRNA enrichment (Xiao and Wilusz 2019). GEO:
198 Gene Expression Omnibus (Barrett et al. 2013), BIGD NGDC: National Genomics Data Center (National
199 Genomics Data Center Members and Partners 2020)

200 *CirComPara2 performs consistently better than single methods*

201 The performance of the methods on the real RNA-seq data agreed with the analysis on simulated data
202 and previous benchmark study results (Zeng et al. 2017; Hansen 2018). Segemehl showed high recall
203 (median 0.75; Figure 4a) but the lowest precision (median 0.92; Figure 4b), whereas C2BW, as
204 expected from annotation-based algorithms, showed the most reliable predictions (median precision
205 0.97; Figure 4b). Similar to the evaluation on the simulated data set, we computed the F_1-scores on the
206 real data set. According to F_1-score medians, CirComPara2 had the highest (0.91) and significantly
207 different value ($q < 0.001$; Figure 4c), and substantially outperformed the next best method, Segemehl
208 (median F_1-score 0.82), with a 0.09 F_1-score difference (Supplementary Table S2). The highest F_1-
209 score, achieved by CirComPara2, resulted from a significantly larger recall (median 0.86; $q < 0.001$;
210 Figure 4a) and a negligible loss of precision compared to the other methods (0.01 median reduction to
211 the most precise method; Figure 4b). Moreover, CirComPara2 had the narrowest interquartile range of
212 F_1-scores across the real-data sets (0.11; Figure 4c; Supplementary Table S2), proving that it is robust
213 and almost unaffected by the experimental scenario. In light of the more challenging nature of real

compared to simulated data (Szabo and Salzman 2016; Chen et al. 2015), these results further confirmed the advantage of CirComPara2 over the methods here assessed.

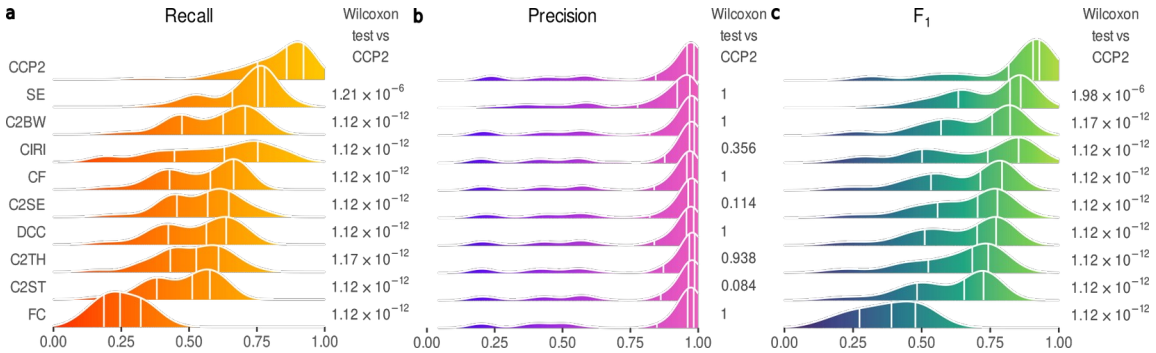

Figure 4. Performance of circRNA detection methods on the real RNA-seq data set. Results of the analysis on the 71 pairs of samples of the real RNA-seq data set. Density plot of (**c**) recall, (**d**) precision and (**e**) F_1-score; distribution quartiles are indicated by white vertical lines. On the right-hand side of each plot, the Bonferroni adjusted p-values of Wilcoxon paired test contrasting each method with CirComPara2. Methods' abbreviations as in Figure 1; CCP2: CirComPara2.

Discussion

In the early days after circRNA discovery, bioinformaticians put significant effort into developing circRNA detection methods with highly precise predictions. To this aim, the current circRNA detection algorithms apply filtering procedures to remove FPs, but that may also reject true circRNAs (Zeng et al. 2017; Chen et al. 2015; Hansen 2018; Hansen et al. 2016), increasing the number of FNs and the risk of overlooking circRNAs of interest, as suggested by our analysis. Moreover, the frequent practice of filtering out low count circRNAs to improve precision may result in the loss of differentially expressed elements in comparative studies, as demonstrated by systematic studies on gene expression (Zhu et al. 2019)(Buratin et al. 2020). Notably, unlike for FPs, experimental validations cannot amend the bias derived by FN errors. These considerations prompted us to consider a method evaluation metric that equally weighted precision and recall, such as the F_1-score.

Interestingly, the evaluation by F_1-score in our simulated data analysis highlighted that some methods with diverging precision and recall, such as CIRI2 and Segemehl (Figure 3a), were equally ranked. Moreover, the real data set analysis showed an ample recall variation of each method across the samples, while precision scores were tighter both between and within methods. These results

13

231 confirmed that choosing a single method for circRNA discovery is problematic since no single best
232 performing method exists. We showed that an approach leveraging the advantages brought by each of
233 the integrated methods allows CirComPara2 to perform better than the single methods consistently in
234 different experimental contexts. Besides, we showed that CirComPara2 predictions are more inclusive
235 and robust than single method ones, even upon low count filtering, indicating that the recovered
236 circRNAs missed by other methods are not merely of low abundance.

237 To our knowledge, CirComPara has been the first bioinformatics tool to combine multiple circRNA
238 detection methods in an automated software pipeline. Since its former implementation, the continuous
239 upgrade of CirComPara evolved into a substantially improved new tool, CirComPara2, with an
240 efficacious method combination strategy. Other computational pipelines that employed multiple
241 circRNA detection tools, such as RAISE (Li et al. 2017a) and circRNAwrap (Li et al. 2019),
242 considered the circRNAs predicted by the union or the intersection of all methods, respectively,
243 without validating which method combination was best. Our data show that those two approaches
244 could be suboptimal as they suffer either low precision or sensitivity, whereas the best tradeoff
245 between precision and recall is achieved with predictions of methods taken pairwise.

246 Recently, large circRNA databases have been compiled using method combination strategies
247 (Vromman et al. 2020). For instance, circAtlas2.0 (Wu et al. 2020) retained circRNAs identified by at
248 least two methods among CIRI2, CIRCexplorer2, Findcirc and DCC; the same combination but
249 replacing DCC with circRNA_finder has been used in CircRic (Ruan et al. 2019). Applying these two
250 approaches to our real-data sets analysis, we observed that they performed better than most of the
251 algorithms except CirComPara2 (Supplementary Table S2), which achieved similar precision but
252 slightly more than 0.2 better recall. Such a result suggests that CirComPara2 could be employed to
253 compile comprehensive and reliable circRNA databases in future works.

254 Importantly, the computational pipelines used to compile circAtlas2.0 and CircRic were not
255 implemented as software tools available to the scientific community. Instead, we made the automated
256 and computationally efficient CirComPara2 pipeline ready-to-use and portable through a Docker

container, freeing bioinformaticians from the several difficulties posed by implementing a computational pipeline, such as installation of multiple tools, software portability, code maintenance and documentation (Menegidio et al. 2018).

CirComPara2 is the only tool that aggregates various method expression estimates into unified values that eliminate redundant counts of BJR identified by multiple tools without relying on additional re-alignment of the reads. CircRNAwrap, RAISE, and CircAtlas2.0 applied a quantification step downstream of circRNA detection to estimate the circRNA expression. They considered the re-alignment of the reads onto pseudo reference circRNA sequences through Sailfish-cir (Li et al. 2017b), RAISE itself, and CIRIquant (Zhang et al. 2020), respectively, thus increasing the computational requirements of the pipeline. Nevertheless, both Sailfish-cir and CIRIquant can be easily applied downstream to CirComPara2 predictions.

CircRNA studies are likely to grow in several branches of biology, both on model (Weigelt et al. 2020) and non-model organisms and beyond the biomedical field (Wu et al. 2021a; Chu et al. 2021; Liang et al. 2019), prompting the development of improved tools allowing more extensive circRNA investigation to unravel circRNA-related condition peculiarities, such as differential circRNA expression (Gaffo et al. 2019; Wu et al. 2021c; Tian et al. 2020), imbalances of the CLP (Buratin et al. 2020), and prevailing circular transcript isoform expression (Izuogu et al. 2018). CirComPara2 is a resource tool meeting these needs, as already proved by the successful application of its embryonic implementations in several studies of human diseases and of other species, including plants (Gaffo et al. 2019; Frydrych Capelari et al. 2019; Buratin et al. 2020; Dal Molin et al. 2020).

In this work, we described the main features of CirComPara2, our automated and computationally efficient software pipeline for circRNA expression characterisation that also allows traditional gene expression analysis and the computation of circRNA to host-gene linear transcript abundance. Importantly, we demonstrated the asset given by the CirComPara2 method combination strategy for circRNA discovery, which provides robust and inclusive predictions in diverse biological contexts. With CirComPara2, we aim to provide a helpful bioinformatics tool to obtain a more comprehensive

15

picture of transcriptomes and boost the understanding of circRNA features, biological and pathogenetic roles.

Methods

Simulated data set

CircRNA reads were simulated with the CIRI_simulator from the CIRI2 tool suite using the whole GRCh38 human genome and Gencode v29 gene annotation.

The parameters used in ciri_simulator were: -C 20 -LC 0 -R 1 -LR 1 -L 101 -E 1 -CHR1 0 -M 250 -M2 450 -PM 0 -S 70 -S2 0 -SE 0 -PSI 10.

A total of 10% (30% for the highly pruned annotation simulation) of the gene annotation for the simulated circRNA parent genes was removed from the annotation file and used as input to circRNA detection methods and to simulate linear transcript reads with Polyester (the reads_per_transcript parameter was set to 300). The linear transcript read files were then concatenated to the circRNA read files.

Code for generating the simulated data is available as a software tool *CCP_simulator* at https://github.com/egaffo/CCP_simulator. The parameter ANNOPARTS used in *CCP_simulator* for the two filters on gene annotation (standard and high pruning) was set to "85,4,5,1,0,0,0" and "60,15,10,2.5,6,3.5,3", respectively.

Real-data sets

Overall, 71 samples with matched rRNA depletion and rRNA depletion followed by RNase R treatment libraries from six studies (Table 1), for a total of 142 samples processed, were retrieved from Gene Expression Omnibus (GEO) or the National Genomics Data Center (NGDC) databases.

Reads from PRJCA000751 were trimmed to 150bp in the preprocessing phase, as reported in the original work.

Method predictions in genomic scaffolds or from the mitochondrial genome were not considered. CircRNAs predicted with a length shorter than the library read length or longer than the longest gene expressed in the sample (computed as genes with TPM ≥ 1 computed by StringTie v2.1.4) were

16

filtered out.

CircRNA detection methods' parameters

The following genome and gene annotations from the Ensembl database were used in the analyses: GRCh38 human genome and v97 gene annotation, Mmul 10 *Macaca mulatta* genome and v101 gene annotation, and GRCm38 *Mus musculus* genome and v101 gene annotation.

CirComPara2 default parameters were set for the analyses, which are as follows:

adaptors from the Trimmomatic v0.39 TruSeq3-PE-2.fa file; PREPROCESSOR = "trimmomatic"; PREPROCESSOR_PARAMS = "MAXINFO:40:0.5 LEADING:20 TRAILING:20 SLIDINGWINDOW:4:30 MINLEN:50 AVGQUAL:26". For the PRJCA000751 data sets, the CROP:150 option was appended to the parameter. STAR_PARAMS = '--runRNGseed 123 --outSJfilterOverhangMin 15 15 15 15 --alignSJoverhangMin 15 --alignSJDBoverhangMin 15 --seedSearchStartLmax 30 --outFilterScoreMin 1 --outFilterMatchNmin 1 --outFilterMismatchNmax 2 --chimSegmentMin 15 --chimScoreMin 15 --chimScoreSeparation 10 --chimJunctionOverhangMin 15'. CIRCRNA_METHODS = 'circexplorer2_bwa, circexplorer2_segemehl, circexplorer2_star, circexplorer2_tophat, ciri, dcc, findcirc' ('circrna_finder' and 'testrealign' values were used in additional runs to obtain CircRNA_finder and Segemhel predictions); CPUS = 12; BWA_PARAMS = '-T 19'; SEGEMEHL_PARAMS = '-D 0'; BOWTIE2_PARAMS = '--reorder --score-min=C,-15,0 -q --seed 123'; DCC_EXTRA_PARAMS = '-fg -M -F -Nr 1 1 -N'; TESTREALIGN_PARAMS = 'q median_1'; FINDCIRC_EXTRA_PARAMS = '--best-qual 40 --filter-tags UNAMBIGUOUS_BP --filter-tags ANCHOR_UNIQUE' (this setting implements the optimization suggested by Hansen (Hansen 2018)). MIN_METHODS = 2; MIN_READS = 2; CIRC_MAPPING = "{'SE': ['STAR','TOPHAT','BOWTIE2'],'PE':['BWA','SEGEMEHL']}"; HISAT2_PARAMS = '--seed 123'.

Segemehl predictions reported in the sngl.bed and trns.txt files were merged to include spliced reads spanning >20,000 bps, as CirComPara2 performs it.

Read alignment and circRNA detection methods' parameters were set as the value in the corresponding CirComPara2 parameters. All other parameters not mentioned were left with default values.

17

336 *Details of the CirComPara2 method*

337 CircRNA expression estimates in CirComPara2 are represented as the sum of all unique back-splice
338 junction reads (BJRs) identified by the circRNA methods. Then, the number of BJR fragments is
339 counted while keeping track of the number of methods detecting each circRNAs. Finally, circRNAs
340 with at least two reads identified by two or more methods are reported. The MIN_METHODS and
341 MIN_READS options can be used to modify the required minimum reads and methods.

342 The CirComPara2 pipeline considers a preliminary alignment step that maps the reads linearly on the
343 genome. Linearly aligned reads are then used to characterise canonical gene and transcript expression
344 and count the linearly spliced reads spanning the back-splice junctions. Instead, linearly unmapped
345 reads are used as input to the circRNA detection methods.

346 *Evaluation metrics and statistical tests*

347 In evaluating method predictions, to compare samples with their matched control libraries, the
348 different sequencing depths of the control library and the biochemical variability of the
349 exoribonuclease in RNase R treatment have been adjusted by considering the proportion between the
350 expression of the predicted circRNAs and the linear transcripts sharing the back-spliced exons. In
351 each sample and for each method, we computed the circular-to-linear expression proportion (CLP) of
352 each predicted circRNA by counting the number of reads back-spliced (BS_{reads}) and linearly-spliced
353 (LS_{reads}) on the circRNA back-splice junctions, respectively. Then, the CLPs were calculated as
354 $BS_{reads}/(BS_{reads}+ LS_{reads})$. Szabo and Salzman (Szabo and Salzman 2016) also suggested using the ratio
355 between the expression of circRNAs and their linear counterparts to overcome the evaluation issues
356 deriving by RNase R treated control samples.

357 Assuming that RNase R should degrade linear transcripts more than circRNAs regardless of its
358 efficiency in different samples, circRNAs with an equal or increased CLP in the control samples were
359 deemed true-positives (TPs); false-positives (FPs) otherwise. Only circRNAs detected in the control
360 libraries were considered to limit incorrect FP calls due to a lower sequencing depth of the control
361 sample and to RNase R sensitive circRNAs (Jeck et al. 2013).

362 The precision was defined as TP/(TP + FP). The recall was computed as TP/(TP + FN) and F_1-score

as (2 × Precision × Recall)/(Precision + Recall), where TP and FN denote the true-positive and false-negative numbers.

Wilcoxon one-tailed paired tests were used to compare CirComPara2 greater recall and F_1-score or lower precision with each method. Bonferroni multiple test correction was applied to compute adjusted p-values (q-values). The q-values reported in the main text refer to the highest value among the pairwise comparisons for recall and F_1-score, whereas to the lowest value for precision comparisons.

Software versions

The software version of the circRNA detection methods and chimeric read aligners used in this study: CIRCexplorer2 (Zhang et al. 2016) v2.3.8, CircRNA_finder (Westholm et al. 2014) v1.1, CIRI2 (Gao et al. 2018) v2.0.6, DCC v0.4.8 (Cheng et al. 2016), Findcirc (Memczak et al. 2013) v1.2, Segemehl (Hoffmann et al. 2014) v0.3.4 (used also by CIRCexplorer2), BWA MEM (Li 2013) v0.7.15 (used by CIRI2 and CIRCexplorer2), STAR (Dobin et al. 2013) v2.6.1e (used by circRNA_finder, DCC, and CIRCexplorer2), TopHat2 (Kim and Salzberg 2011) v2.1.0 (used for the TopHat-Fusion algorithm by CIRCexplorer2). CirComPara2 v0.1.2.1 was used to run the analysis.

Data access

All the analysed data have been previously published and the accessions are referenced in Table 1.

Competing interests

The authors declare that they have no competing interests.

Acknowledgements

We thank Dr Geertruij te Kronnie for insightful suggestions and critical revision of the manuscript.

Funding

This work has been supported by the Italian Ministry of Education, Universities, and Research grant PRIN 2017 2017PPS2X4_003 (S.B.), AIRC, Milan, Italy Investigator Grant 2017 20052 (S.B.), and Fondazione Umberto Veronesi, Milan, Italy Fellowship 2020 (E.G.).

Author Contributions

Conceptualisation, E.G.; Data curation, E.G.; Formal Analysis, E.G. and A.B.; Funding Acquisition, E.G. and S.B.; Methodology, E.G. and A.B.; Project Administration, S.B.; Resources, S.B.; Software, E.G. and A.B.; Supervision, E.G. and S.B.; Visualization, E.G.; Writing – Original Draft, E.G., A.B. and S.B.; Writing – Review & Editing, E.G., A.B., A.D.M. and S.B. All authors read and approved the final manuscript.

References

Andrews S. 2010. FastQC: a quality control tool for high throughput sequence data. *FastQC*. http://www.bioinformatics.babraham.ac.uk/projects/fastqc (Accessed May 5, 2021).

Awan FM, Yang BB, Naz A, Hanif A, Ikram A, Obaid A, Malik A, Janjua HA, Ali A, Sharif S. 2021. The emerging role and significance of circular RNAs in viral infections and antiviral immune responses: possible implication as theranostic agents. *RNA Biol* **18**: 1–15.

Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, Marshall KA, Phillippy KH, Sherman PM, Holko M, et al. 2013. NCBI GEO: archive for functional genomics data sets--update. *Nucleic Acids Res* **41**: D991–5.

Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**: 2114–2120. http://dx.doi.org/10.1093/bioinformatics/btu170.

Bonizzato A, Gaffo E, Te Kronnie G, Bortoluzzi S. 2016. CircRNAs in hematopoiesis and hematological malignancies. *Blood Cancer J* **6**: e483.

Buratin A, Paganin M, Gaffo E, Dal Molin A, Roels J, Germano G, Siddi MT, Serafin V, De Decker M, Gachet S, et al. 2020. Large-scale circular RNA deregulation in T-ALL: unlocking unique ectopic expression of molecular subtypes. *Blood Adv* **4**: 5902–5914.

Cheng J, Metge F, Dieterich C. 2016. Specific identification and quantification of circular RNAs from sequencing data. *Bioinformatics* **32**: 1094–1096.

20

412 Chen I, Chen C-Y, Chuang T-J. 2015. Biogenesis, identification, and function of exonic circular
413 RNAs. *Wiley Interdiscip Rev RNA* **6**: 563–579.

414 Chen L, Wang C, Sun H, Wang J, Liang Y, Wang Y, Wong G. 2020. The bioinformatics toolbox for
415 circRNA discovery and analysis. *Brief Bioinform*. http://dx.doi.org/10.1093/bib/bbaa001.

416 Chen S, Huang V, Xu X, Livingstone J, Soares F, Jeon J, Zeng Y, Hua JT, Petricca J, Guo H, et al.
417 2019. Widespread and Functional RNA Circularization in Localized Prostate Cancer. *Cell* **176**:
418 831–843.e22.

419 Chen Y, Lun ATL, Smyth GK. 2016. From reads to genes to pathways: differential expression
420 analysis of RNA-Seq experiments using Rsubread and the edgeR quasi-likelihood pipeline.
421 *F1000Res* **5**: 1438.

422 Chu Q, Zheng W, Su H, Zhang L, Chang R, Gao W, Xu T. 2021. A Highly Conserved Circular RNA
423 circRasGEF1B Enhances Antiviral Immunity by Regulating miR-21-3p/MITA Pathway in
424 Lower Vertebrates. *J Virol*. http://dx.doi.org/10.1128/JVI.02145-20.

425 Dal Molin A, Hofmans M, Gaffo E, Buratin A, Cavé H, Flotho C, de Haas V, Niemeyer CM, Stary J,
426 Van Vlierberghe P, et al. 2020. CircRNAs Dysregulated in Juvenile Myelomonocytic Leukemia:
427 CircMCTP1 Stands Out. *Front Cell Dev Biol* **8**: 613540.

428 Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR.
429 2013. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**: 15–21.

430 Du WW, Yang W, Liu E, Yang Z, Dhaliwal P, Yang BB. 2016. Foxo3 circular RNA retards cell cycle
431 progression via forming ternary complexes with p21 and CDK2. *Nucleic Acids Res* **44**: 2846–
432 2858.

433 Frydrych Capelari É, da Fonseca GC, Guzman F, Margis R. 2019. Circular and Micro RNAs from
434 Flowers Are Simultaneously Isolated from AGO-IP Libraries. *Plants* **8**.
435 http://dx.doi.org/10.3390/plants8090302.

436 Gaffo E, Boldrin E, Dal Molin A, Bresolin S, Bonizzato A, Trentin L, Frasson C, Debatin K-M,
437 Meyer LH, Te Kronnie G, et al. 2019. Circular RNA differential expression in blood cell
438 populations and exploration of circRNA deregulation in pediatric acute lymphoblastic leukemia.
439 *Sci Rep* **9**: 14670.

440 Gaffo E, Bonizzato A, Kronnie GT, Bortoluzzi S. 2017. CirComPara: A Multi-Method Comparative
441 Bioinformatics Pipeline to Detect and Study circRNAs from RNA-seq Data. *Noncoding RNA* **3**.
442 http://dx.doi.org/10.3390/ncrna3010008.

443 Gao Y, Wang J, Zheng Y, Zhang J, Chen S, Zhao F. 2016. Comprehensive identification of internal
444 structure and alternative splicing events in circular RNAs. *Nat Commun* **7**: 12060.

445 Gao Y, Zhang J, Zhao F. 2018. Circular RNA identification based on multiple seed matching. *Brief*
446 *Bioinform* **19**: 803–810.

447 Hanniford D, Ulloa-Morales A, Karz A, Berzoti-Coelho MG, Moubarak RS, Sánchez-Sendra B,
448 Kloetgen A, Davalos V, Imig J, Wu P, et al. 2020. Epigenetic Silencing of CDR1as Drives
449 IGF2BP3-Mediated Melanoma Invasion and Metastasis. *Cancer Cell* **37**: 55–70.e15.

450 Hansen TB. 2018. Improved circRNA Identification by Combining Prediction Algorithms. *Front Cell*
451 *Dev Biol* **6**: 20.

452 Hansen TB, Jensen TI, Clausen BH, Bramsen JB, Finsen B, Damgaard CK, Kjems J. 2013. Natural
453 RNA circles function as efficient microRNA sponges. *Nature* **495**: 384–388.

454 Hansen TB, Venø MT, Damgaard CK, Kjems J. 2016. Comparison of circular RNA prediction tools.
455 *Nucleic Acids Res* **44**: e58.

456 Hoffmann S, Otto C, Doose G, Tanzer A, Langenberger D, Christ S, Kunz M, Holdt LM, Teupser D,
457 Hackermüller J, et al. 2014. A multi-split mapping algorithm for circular RNA, splicing, trans-
458 splicing and fusion detection. *Genome Biol* **15**: R34.

459 Hua JT, Chen S, He HH. 2019. Landscape of Noncoding RNA in Prostate Cancer. *Trends Genet* **35**:

460 840–851.

461 Izuogu OG, Alhasan AA, Mellough C, Collin J, Gallon R, Hyslop J, Mastrorosa FK, Ehrmann I, Lako
462 M, Elliott DJ, et al. 2018. Analysis of human ES cell differentiation establishes that the dominant
463 isoforms of the lncRNAs RMST and FIRRE are circular. *BMC Genomics* **19**: 276.

464 Jeck WR, Sorrentino JA, Wang K, Slevin MK, Burd CE, Liu J, Marzluff WF, Sharpless NE. 2013.
465 Circular RNAs are abundant, conserved, and associated with ALU repeats. *RNA* **19**: 141–157.

466 Ji P, Wu W, Chen S, Zheng Y, Zhou L, Zhang J, Cheng H, Yan J, Zhang S, Yang P, et al. 2019.
467 Expanded Expression Landscape and Prioritization of Circular RNAs in Mammals. *Cell Rep* **26**:
468 3444–3460.e5.

469 Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. 2019. Graph-based genome alignment and
470 genotyping with HISAT2 and HISAT-genotype. *Nat Biotechnol* **37**: 907–915.

471 Kim D, Salzberg SL. 2011. TopHat-Fusion: an algorithm for discovery of novel fusion transcripts.
472 *Genome Biol* **12**: R72.

473 Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**: 357–
474 359.

475 Liang Y, Zhang Y, Xu L, Zhou D, Jin Z, Zhou H, Lin S, Cao J, Huang L. 2019. CircRNA Expression
476 Pattern and ceRNA and miRNA–mRNA Networks Involved in Anther Development in the CMS
477 Line of Brassica campestris. *International Journal of Molecular Sciences* **20**: 4808.
478 http://dx.doi.org/10.3390/ijms20194808.

479 Li H. 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM.
480 http://arxiv.org/abs/1303.3997 (Accessed October 7, 2020).

481 Li L, Bu D, Zhao Y. 2019. CircRNAwrap - a flexible pipeline for circRNA identification, transcript
482 prediction, and abundance estimation. *FEBS Lett* **593**: 1179–1189.

483 Li L, Zheng Y-C, Kayani MUR, Xu W, Wang G-Q, Sun P, Ao N, Zhang L-N, Gu Z-Q, Wu L-C, et al.
484 2017a. Comprehensive analysis of circRNA expression profiles in humans by RAISE. *Int J*
485 *Oncol* **51**: 1625–1638.

486 Li M, Xie X, Zhou J, Sheng M, Yin X, Ko E-A, Zhou T, Gu W. 2017b. Quantifying circular RNA
487 expression from RNA-seq data using model-based framework. *Bioinformatics* **33**: 2131–2139.

488 Li Z, Huang C, Bao C, Chen L, Lin M, Wang X, Zhong G, Yu B, Hu W, Dai L, et al. 2015. Exon-
489 intron circular RNAs regulate transcription in the nucleus. *Nat Struct Mol Biol* **22**: 256–264.

490 Love MI, Soneson C, Hickey PF, Johnson LK, Pierce NT, Shepherd L, Morgan M, Patro R. 2020.
491 Tximeta: Reference sequence checksums for provenance identification in RNA-seq. *PLoS*
492 *Comput Biol* **16**: e1007664.

493 Memczak S, Jens M, Elefsinioti A, Torti F, Krueger J, Rybak A, Maier L, Mackowiak SD, Gregersen
494 LH, Munschauer M, et al. 2013. Circular RNAs are a large class of animal RNAs with regulatory
495 potency. *Nature* **495**: 333–338.

496 Menegidio FB, Jabes DL, Costa de Oliveira R, Nunes LR. 2018. Dugong: a Docker image, based on
497 Ubuntu Linux, focused on reproducibility and replicability for bioinformatics analyses.
498 *Bioinformatics* **34**: 514–515.

499 National Genomics Data Center Members and Partners. 2020. Database Resources of the National
500 Genomics Data Center in 2020. *Nucleic Acids Res* **48**: D24–D33.

501 Pertea M, Pertea GM, Antonescu CM, Chang T-C, Mendell JT, Salzberg SL. 2015. StringTie enables
502 improved reconstruction of a transcriptome from RNA-seq reads. *Nature Biotechnology* **33**: 290–
503 295. http://dx.doi.org/10.1038/nbt.3122.

504 Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features.
505 *Bioinformatics* **26**: 841–842. http://dx.doi.org/10.1093/bioinformatics/btq033.

506 Rajappa A, Banerjee S, Sharma V, Khandelia P. 2020. Circular RNAs: Emerging Role in Cancer

507 Diagnostics and Therapeutics. *Front Mol Biosci* **7**: 577938.

508 Ruan H, Xiang Y, Ko J, Li S, Jing Y, Zhu X, Ye Y, Zhang Z, Mills T, Feng J, et al. 2019.
509 Comprehensive characterization of circular RNAs in ~ 1000 human cancer cell lines. *Genome*
510 *Med* **11**: 55.

511 Santer L, Bär C, Thum T. 2019. Circular RNAs: A Novel Class of Functional RNA Molecules with a
512 Therapeutic Perspective. *Mol Ther* **27**: 1350–1363.

513 Slack FJ, Chinnaiyan AM. 2019. The Role of Non-coding RNAs in Oncology. *Cell* **179**: 1033–1055.
514 http://dx.doi.org/10.1016/j.cell.2019.10.017.

515 Soneson C, Love MI, Robinson MD. 2015. Differential analyses for RNA-seq: transcript-level
516 estimates improve gene-level inferences. *F1000Res* **4**: 1521.

517 Szabo L, Salzman J. 2016. Detecting circular RNAs: bioinformatic and experimental challenges.
518 *Nature Reviews Genetics* **17**: 679–692. http://dx.doi.org/10.1038/nrg.2016.114.

519 Tange O. 2020. GNU Parallel 20200922 (’Ginsburg'). https://zenodo.org/record/4045386 (Accessed
520 April 23, 2021).

521 Tian J, Fu Y, Li Q, Xu Y, Xi X, Zheng Y, Yu L, Wang Z, Yu B, Tian J. 2020. Differential Expression
522 and Bioinformatics Analysis of CircRNA in PDGF-BB-Induced Vascular Smooth Muscle Cells.
523 *Front Genet* **11**: 530.

524 Vromman M, Vandesompele J, Volders P-J. 2020. Closing the circle: current state and perspectives of
525 circular RNA databases. *Brief Bioinform*. http://dx.doi.org/10.1093/bib/bbz175.

526 Weigelt CM, Sehgal R, Tain LS, Cheng J, Eßer J, Pahl A, Dieterich C, Grönke S, Partridge L. 2020.
527 An Insulin-Sensitive Circular RNA that Regulates Lifespan in Drosophila. *Mol Cell* **79**: 268–
528 279.e5.

529 Westholm JO, Miura P, Olson S, Shenker S, Joseph B, Sanfilippo P, Celniker SE, Graveley BR, Lai

530 EC. 2014. Genome-wide analysis of drosophila circular RNAs reveals their structural and
531 sequence properties and age-dependent neural accumulation. *Cell Rep* **9**: 1966–1980.

532 Wu Q, Ning X, Sun L. 2021a. Megalocytivirus Induces Complicated Fish Immune Response at
533 Multiple RNA Levels Involving mRNA, miRNA, and circRNA. *Int J Mol Sci* **22**.
534 http://dx.doi.org/10.3390/ijms22063156.

535 Wu W, Ji P, Zhao F. 2020. CircAtlas: an integrated resource of one million highly accurate circular
536 RNAs from 1070 vertebrate transcriptomes. *Genome Biol* **21**: 101.

537 Wu X, Xiao S, Zhang M, Yang L, Zhong J, Li B, Li F, Xia X, Li X, Zhou H, et al. 2021b. A novel
538 protein encoded by circular SMO RNA is essential for Hedgehog signaling activation and
539 glioblastoma tumorigenicity. *Genome Biol* **22**: 33.

540 Wu Y, Zhao T, Deng R, Xia X, Li B, Wang X. 2021c. A study of differential circRNA and lncRNA
541 expressions in COVID-19-infected peripheral blood. *Sci Rep* **11**: 7991.

542 Xiao M-S, Ai Y, Wilusz JE. 2020. Biogenesis and Functions of Circular RNAs Come into Focus.
543 *Trends Cell Biol* **30**: 226–240.

544 Xiao M-S, Wilusz JE. 2019. An improved method for circular RNA purification using RNase R that
545 efficiently removes linear RNAs containing G-quadruplexes or structured 3' ends. *Nucleic Acids*
546 *Res* **47**: 8755–8769.

547 Zeng X, Lin W, Guo M, Zou Q. 2017. A comprehensive overview and evaluation of circular RNA
548 detection tools. *PLoS Comput Biol* **13**: e1005420.

549 Zhang J, Chen S, Yang J, Zhao F. 2020. Accurate quantification of circular RNAs identifies extensive
550 circular isoform switching events. *Nat Commun* **11**: 90.

551 Zhang X-O, Dong R, Zhang Y, Zhang J-L, Luo Z, Zhang J, Chen L-L, Yang L. 2016. Diverse
552 alternative back-splicing and alternative splicing landscape of circular RNAs. *Genome Res* **26**:
553 1277–1287.

554 Zhang X-O, Wang H-B, Zhang Y, Lu X, Chen L-L, Yang L. 2014. Complementary sequence-
555 mediated exon circularization. *Cell* **159**: 134–147.

556 Zhu A, Ibrahim JG, Love MI. 2019. Heavy-tailed prior distributions for sequence count data:
557 removing the noise and preserving large differences. *Bioinformatics* **35**: 2084–2092.

27