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.)

Keywords: circRNAs, bioinformatics, transcriptomics, RNA-seq

Abstract

Current methods for the identification of circular RNAs (circRNAs) suffer from low discovery rates and inconsistent performance in diverse data sets. Therefore, the applied detection algorithm can bias high-throughput study findings by missing relevant circRNAs. Here, we show that our bioinformatics tool CirComPara2 (https://github.com/egaffo/CirComPara2), by combining multiple circRNA detection methods, consistently achieves high recall rates without loss of precision in simulated and different real data sets.

Main

Recent research uncovered that eukaryotic transcriptomes comprise thousands of stable circular RNAs (circRNAs), which can modulate cellular processes by encoding functional peptides, and by interacting with microRNA and RNA-binding proteins [1]. The regulatory role of circRNAs and their potential as biomarkers actuate the integration of circRNA in conventional transcriptome analysis, especially in cancer research and studies of pathological conditions.

CircRNAs originate by transcript back-splicing, a process where the molecule's 3' and 5' ends are covalently joined. Studies of circRNAs rapidly increased in pace with the development of bioinformatics approaches that identify the sequences spanning circRNA back-splice junctions from total RNA-seq data. To date, several methods for circRNA identification have been developed [2], but none clearly outperforms the others since they provide either highly sensitive or highly precise predictions [3–6]. To improve precision, circRNA detection algorithms employ filtering procedures [3–6] that remove false-positive guesses (FPs), but that may also reject true circRNAs, thus increasing the number of false-negative predictions (FNs). Differently from FPs, FN errors cannot be amended by experimental validation and can result in overlooking circRNAs of interest, as suggested by the following analysis.

We analysed simulated RNA-seq data of 5,680 circRNAs from the whole human genome with nine circRNA discovery pipelines: circRNA_finder [7], CIRI2 [8], DCC [9], Findcirc [10], Segemehl [11], and CIRCexplorer2 [12] coupled to either BWA [13], Segemehl, STAR [14] and TopHat-Fusion [15]. On average,

38 49% of each method's FNs were above the overall circRNA median expression (Figure 1a), showing
39 that nearly half of the missed circRNAs had a considerable expression. The output of the methods was
40 further examined for 1,945 circRNAs that were missed by one or more tools (the FN set) by
41 inspecting each algorithm for correct identification of the circRNAs in the FN set. Only 4% circRNAs
42 in the FN set were not discovered by any of the nine methods (Figure 1b), suggesting that the
43 application of multiple algorithms with dissimilar and possibly complementary features can improve
44 the circRNA detection rate.
45 However, combining predictions from different methods can worsen the overall precision if the
46 number of FPs is not controlled. Our former method CirComPara [16] combined four circRNA detection
47 tools and limited the number of FPs by excluding the circRNAs identified by single methods,
48 considering that circRNAs predicted by two or more algorithms helps to discard many FPs [5].
49 Interestingly, 95% of the FNs in our simulation were detected by two or more methods, with distinct
50 circRNA sets identified by different method combinations, and only 1% was identified by single tools
51 (Segemehl, CIRCexplorer2 on BWA, and CIRCexplorer2 on TopHat-Fusion; Figure 1b). This
52 observation prompted us to implement CirComPara2, which integrates the nine circRNA detection
53 algorithms presented above. Although circRNA_finder and Segemehl can be enabled in CirComPara2
54 by custom parameters, they are not considered in the default setting (Figure 1c) since in the simulated
55 data analysis a high number of FPs affected the conjoint predictions of methods combinations that
56 included circRNA_finder or Segemehl (Supplementary Figure S1). This agreed with the previous
57 observation that certain method combinations do not ameliorate the discovery of true circRNAs [5].
58 Besides, we assessed whether CirComPara2 achieved the optimal balance between the numbers of
59 recovered FNs and introduced FPs by using the conjoint predictions of at least two methods or by a
60 different number of methods combined. The predictions based on all methods resulted in the highest
61 recall (0.99), but at the cost of the lowest precision (0.90; Figure 1d), indicating that the gained FNs
62 were outweighed by the accumulation of FPs. Next, subsetting predictions to the circRNAs commonly
63 detected by at least two tools showed a slightly reduced recall (0.98) with a substantially increased
64 precision (0.99). Instead, selecting the conjoint predictions of three or more methods largely reduced
65 the recall (0.96 to 0.68) with only a modest increase in precision (0.99 to 1.00). To rank the best

performing method combinations, we calculated F_1-scores, which provide single-value estimates accounting for and evenly weighing both recall and precision. The highest F_1-score achieved by the conjoint predictions of two or more methods (0.99; Figure 1d) indicated and confirmed the optimal trade-off between recall and precision. Therefore, CirComPara2 was set to report the predictions common to at least two methods.

Next, the performances on the simulated data of CirComPara2 and the other nine methods were compared. CirComPara2 obtained the highest F_1-score (0.99) by achieving the highest recall (0.98) while holding a precision comparable to the other algorithms (0.99 versus 0.92 to 1.00; Figure 2a), confirming that CirComPara2 rectified the true circRNAs missed by the other methods. CirComPara2 maintained the highest F_1-score (0.98) even with a largely incomplete gene annotation input (see Methods), which for the annotation-based methods caused a dramatic reduction (up to -0.14) of the recall (and of F_1-scores) (Supplementary Figure S2), indicating that CirComPara2 can leverage the integrated annotation-independent algorithms, and suggesting that CirComPara2 can be efficient also when applied to RNA-seq data of organisms with poor genome annotation.

CircRNA expression analysis usually involves downstream data cleaning to remove background noise signals or mapping errors, and poorly expressed circRNAs of little interest. 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 on the expression abundance. Accordingly, all methods scored higher precision when BJR threshold filters were increased, but with a corresponding reduced recall, indicating that true circRNAs (with low expression) were discarded as well (Supplementary Figure S3). CirComPara2 showed similar behaviour, but maintained the highest recall compared to other methods at equal precision (Figure 2b; Supplementary Table S1), suggesting that the circRNAs recovered by CirComPara2 were true findings of considerable abundance, which may represent relevant circRNAs in real experiments.

To evaluate circRNA detection methods on real RNA-seq data the gold standard is to enrich circRNAs fraction in the sequencing library by treatment with RNase R [3,5,6,8,17], exploiting the circRNA resistance to degradation by the lack of a single-stranded 3′

overhang. CirComPara2 was then assessed on 142 public real RNA-seq data sets from six independent studies including human cell lines, and various human, Rhesus macaque, and mice tissues (Supplementary Table S2). Each sample matched ribosomal RNA-depleted and RNase R-treated libraries, which we used as control samples (see Methods).

Performance of the methods agreed with the simulation analysis and previous benchmark study results [3,5]: Segemehl showed high recall (median 0.75; Figure 2c) but the lowest precision (median 0.92), and CIRCexplorer2 on BWA showed the most reliable predictions (median precision 0.97; Figure 2d) as expected from annotation-based algorithms. Ranking the tools by their median F_1-score on the real data sets, CirComPara2 had the highest (0.91) and significantly different value ($q < 0.001$; Figure 2e), substantially outperforming the method with next best F_1-score, Segemehl (median 0.82; Supplementary Table S3). CirComPara2's highest F_1-score resulted from its significantly larger recall (median 0.86; $q < 0.001$; Figure 2c) and a negligible loss of precision compared to the other methods (0.01 median reduction from the most precise method; Figure 2d). Notably, CirComPara2 had the narrowest interquartile range for F_1-scores across the real data sets (0.11; Figure 2e; Supplementary Table S3), proving robustness in different scenarios. Considering the more challenging nature of real compared to simulated data [4,17], these results confirmed the advantage of CirComPara2 over other methods.

CirComPara2 implements a fully automated computational pipeline for circRNA detection and quantification, from read pre-processing to circRNA annotation (Supplementary Figure S4), with a modular and highly parallelised implementation that make it computationally efficient and resilient. CirComPara2 is available as standalone software and as a Docker image that facilitates installation, portability, and reproducibility.

Recently, large circRNA databases have been compiled using approaches similar to CirComPara2's strategy [18]. For instance, circAtlas2.0 [19] retained circRNAs identified by at least two methods among CIRI2, CIRCexplorer2, Findcirc and DCC; the same combination but replacing DCC with circRNA_finder has been used in CircRic [20]. On our real data sets, these two method combinations performed better than most of the algorithms except CirComPara2 (Supplementary Table S3), which achieved similar precision and a slightly more than 0.2 better recall, proving the asset given by the

5

larger number of tools combined in CirComPara2. In light of the considerations above, CirComPara2 could provide comprehensive and reliable results to compile circRNA databases.

In conclusion, CirComPara2 allows comprehensive investigation of circRNA expression by enhancing the detection of reliable circRNAs compared to state-of-the-art tools and mitigates the biases affecting single-methods, providing robust circRNA predictions in different biological contexts and species.

6

Materials and 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 (Supplementary Table S2), for a total of 142 samples processed, were retrieved from Gene Expression Omnibus (GEO) or the National Genomics Data Center (NGDC) databases.

True positive circRNAs were defined as those having higher circular-to-linear proportion in the matched circRNA enriched sample. To avoid detection bias caused by different sample sequencing depth, only circRNAs detected in the RNase R treated samples were considered.

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.

7

154 CircRNAs predicted with a length shorter than the library read length or longer than the longest gene
155 expressed in the sample (computed as genes with TPM ≥ 1 computed by StringTie v2.1.4) were
156 filtered out.

157 *CircRNA detection methods' parameters*

158 The following genome and gene annotations from the Ensembl database were used in the analyses:
159 GRCh38 human genome and v97 gene annotation, Mmul 10 *Macaca mulatta* genome and v101 gene
160 annotation, and GRCm38 *Mus musculus* genome and v101 gene annotation.
161 CirComPara2 default parameters were set for the analyses, which are as follows:
162 adaptors from the Trimmomatic v0.39 TruSeq3-PE-2.fa file; PREPROCESSOR = "trimmomatic";
163 PREPROCESSOR_PARAMS = "MAXINFO:40:0.5 LEADING:20 TRAILING:20
164 SLIDINGWINDOW:4:30 MINLEN:50 AVGQUAL:26". For the PRJCA000751 data sets, the
165 CROP:150 option was appended to the parameter. STAR_PARAMS = '--runRNGseed 123 --
166 outSJfilterOverhangMin 15 15 15 15 --alignSJoverhangMin 15 --alignSJDBoverhangMin 15 --
167 seedSearchStartLmax 30 --outFilterScoreMin 1 --outFilterMatchNmin 1 --outFilterMismatchNmax 2
168 --chimSegmentMin 15 --chimScoreMin 15 --chimScoreSeparation 10 --chimJunctionOverhangMin
169 15'. CIRCRNA_METHODS = 'circexplorer2_bwa, circexplorer2_segemehl, circexplorer2_star,
170 circexplorer2_tophat, ciri, dcc, findcirc' ('circrna_finder' and 'testrealign' values were used in
171 additional runs to obtain CircRNA_finder and Segemhel predictions); CPUS = 12; BWA_PARAMS =
172 '-T 19'; SEGEMEHL_PARAMS = '-D 0'; BOWTIE2_PARAMS = '--reorder --score-min=C,-15,0 -q
173 --seed 123'; DCC_EXTRA_PARAMS = '-fg -M -F -Nr 1 1 -N'; TESTREALIGN_PARAMS = 'q
174 median_1'; FINDCIRC_EXTRA_PARAMS = '--best-qual 40 --filter-tags UNAMBIGUOUS_BP --
175 filter-tags ANCHOR_UNIQUE' (this setting implements the optimization suggested by Hansen [5]).
176 MIN_METHODS = 2; MIN_READS = 2; CIRC_MAPPING = "{'SE':
177 ['STAR','TOPHAT','BOWTIE2'],'PE':['BWA','SEGEMEHL']}"; HISAT2_PARAMS = '--seed 123'.
178 Segemehl predictions reported in the sngl.bed and trns.txt files were merged to include spliced reads
179 spanning >20,000 bps, as it is performed by CirComPara2.
180 Read alignment and circRNA detection methods' parameters were set as the value in the

181 corresponding CirComPara2 parameters. All other parameters not mentioned were left with default
182 values.

183 *Details of the CirComPara2 method*

184 CircRNA expression estimates in CirComPara2 are represented as the sum of all unique back-splice
185 junction reads (BJRs) identified by the circRNA methods. Then, the number of BJR fragments is
186 counted keeping track of the number of methods detecting each circRNAs. Finally, circRNAs with at
187 least two reads identified by two or more methods are reported. The MIN_METHODS and
188 MIN_READS options can be used to modify the required minimum reads and methods.
189 The CirComPara2 pipeline considers a preliminary alignment step that maps the reads linearly on the
190 genome. Linearly aligned reads are then used for canonical gene and transcript expression
191 characterisation, and to count the linearly spliced reads spanning the back-splice junctions. Instead,
192 linearly unmapped reads are used as input to the circRNA detection methods.

193 *Evaluation metrics and statistical tests*

194 To compare samples with their matched control libraries for evaluating method predictions, different
195 sequencing depths of the control library and the biochemical variability of the exoribonuclease in
196 RNase R treatment have been considered by examining the proportion between the expression of the
197 predicted circRNAs and the linear transcripts sharing the back-spliced exons [17]. In each sample and
198 for each method, we computed the circular-to-linear expression proportion (CLP) of each predicted
199 circRNA by counting the number of reads back-spliced (BS_{reads}) and linearly-spliced (LS_{reads}) on the
200 circRNA back-splice junctions, respectively. Then, the CLPs were calculated as $BS_{reads}/(BS_{reads}+$
201 $LS_{reads})$. Using the ratio between the expression of cirRNAs and their linear counterparts to overcome
202 the evaluation issues deriving by the use of RNase R treated control samples was suggested also by
203 Szabo and Salzman [17].
204 Assuming that RNase R should degrade linear transcripts more than circRNAs regardless of its
205 efficiency in different samples, circRNAs with an equal or increased CLP in the control samples were
206 deemed true-positives (TPs); false-positives (FPs) otherwise. Only circRNAs detected in the control
207 libraries were considered to limit incorrect FP calls due to a lower sequencing depth of the control

208 sample.

209 The precision was defined as TP/(TP + FP). The recall was computed as TP/(TP + FN) and F_1-score
210 as (2 × Precision × Recall)/(Precision + Recall), where TP and FN denote the true-positive and false-
211 negative numbers.

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

217 *Software versions*

218 The software version of the circRNA detection methods and chimeric read aligners used in this study:
219 CIRCexplorer2 [12] v2.3.8, CircRNA_finder [7] v1.1, CIRI2 [8] v2.0.6, DCC v0.4.8 [9], Findcirc [10] v1.2,
220 Segemehl [11] v0.3.4 (used also by CIRCexplorer2), BWA MEM [13] v0.7.15 (used by CIRI2 and
221 CIRCexplorer2), STAR [14] v2.6.1e (used by circRNA_finder, DCC, and CIRCexplorer2), TopHat2 [15]
222 v2.1.0 (used for the TopHat-Fusion algorithm by CIRCexplorer2). CirComPara2 v0.1 was used to run
223 the analysis.

224 **Key Points**

225 - Current circRNA detection methods achieve either high precision or high recall
226 - Current circRNA detection methods do not perform consistently across different data sets
227 - Current circRNA detection methods can overlook circRNAs of interest
228 - CirComPara2 achieves high detection recall with no loss of precision by combining seven
229 methods
230 - CirComPara2 consistently outperforms other methods regardless of the biological context

231 **Funding**

232 This work has been supported by Italian Ministry of Education, Universities, and Research grant
233 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.).

Acknowledgements

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

Author Contributions

Conceptualization, 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 have read and agreed to the published version of the manuscript.

Conflicts of Interest

The authors declare no conflict of interest.

11

245 **Figures**

246 *Figure 1*

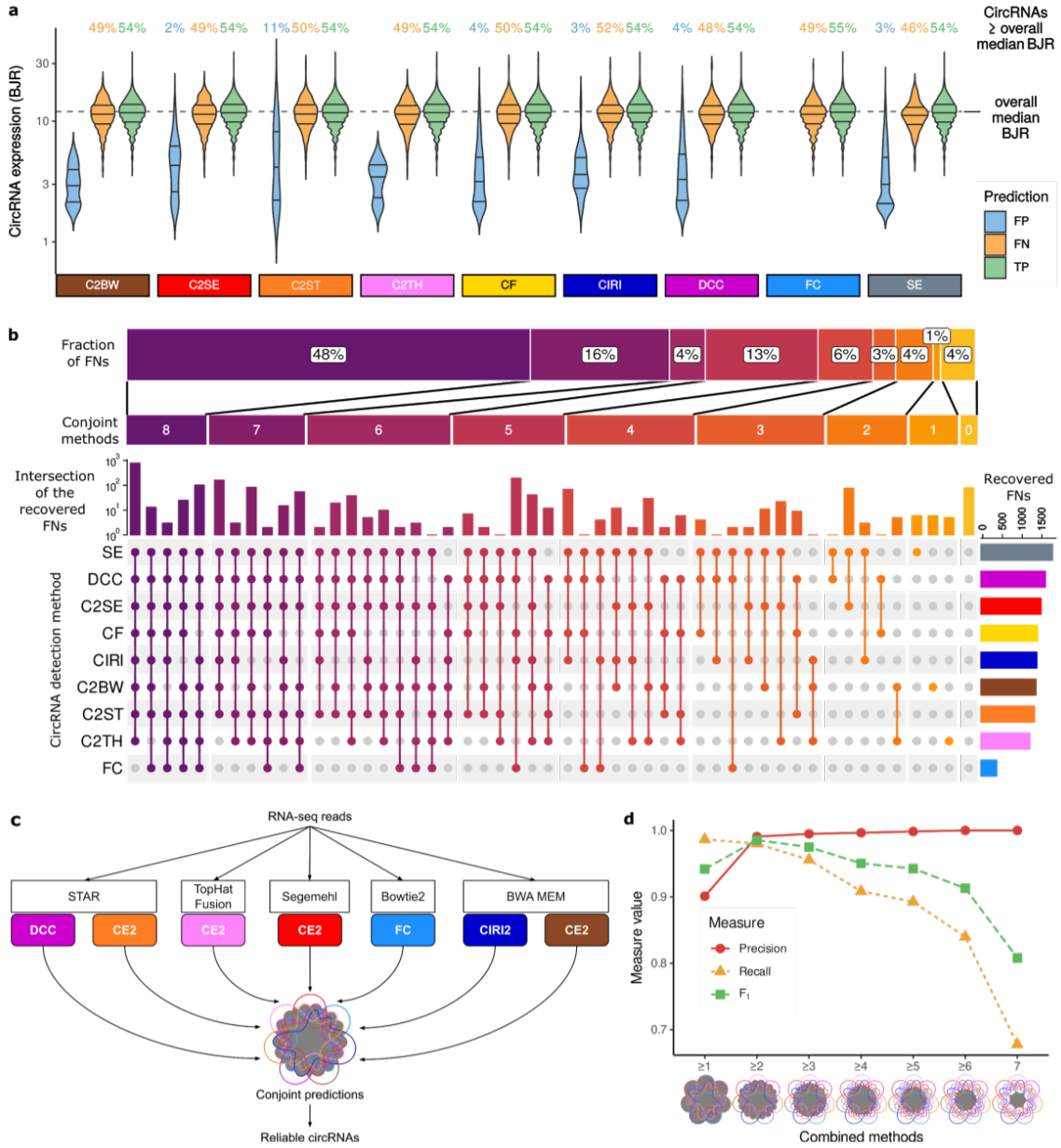

Figure 1. CircRNAs overlooked by common detection methods and the CirComPara2 approach. a, The expression level of predicted circRNAs. BJR: back-splice junction read counts; C2BW: CIRCexplorer2 on BWA; C2SE: CIRCexplorer2 on Segemehl; C2ST: CIRCexplorer2 on STAR; C2TH: CIRCexplorer2 on TopHat-Fusion; 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.

12

Grey dots indicate the methods failing to detect the circRNAs considered in the barchart 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. **c**, Workflow of the CirComPara2 strategy showing circRNA detection methods (coloured rounded corner boxes) with the respective chimeric read aligners (white boxes). The central Venn diagram represents the optimized combination of methods' prediction intersections implemented by CirComPara2 with circRNAs conjointly detected by two or more methods (grey filled intersections) retained as default. **d**, 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.

13

247 *Figure 2*

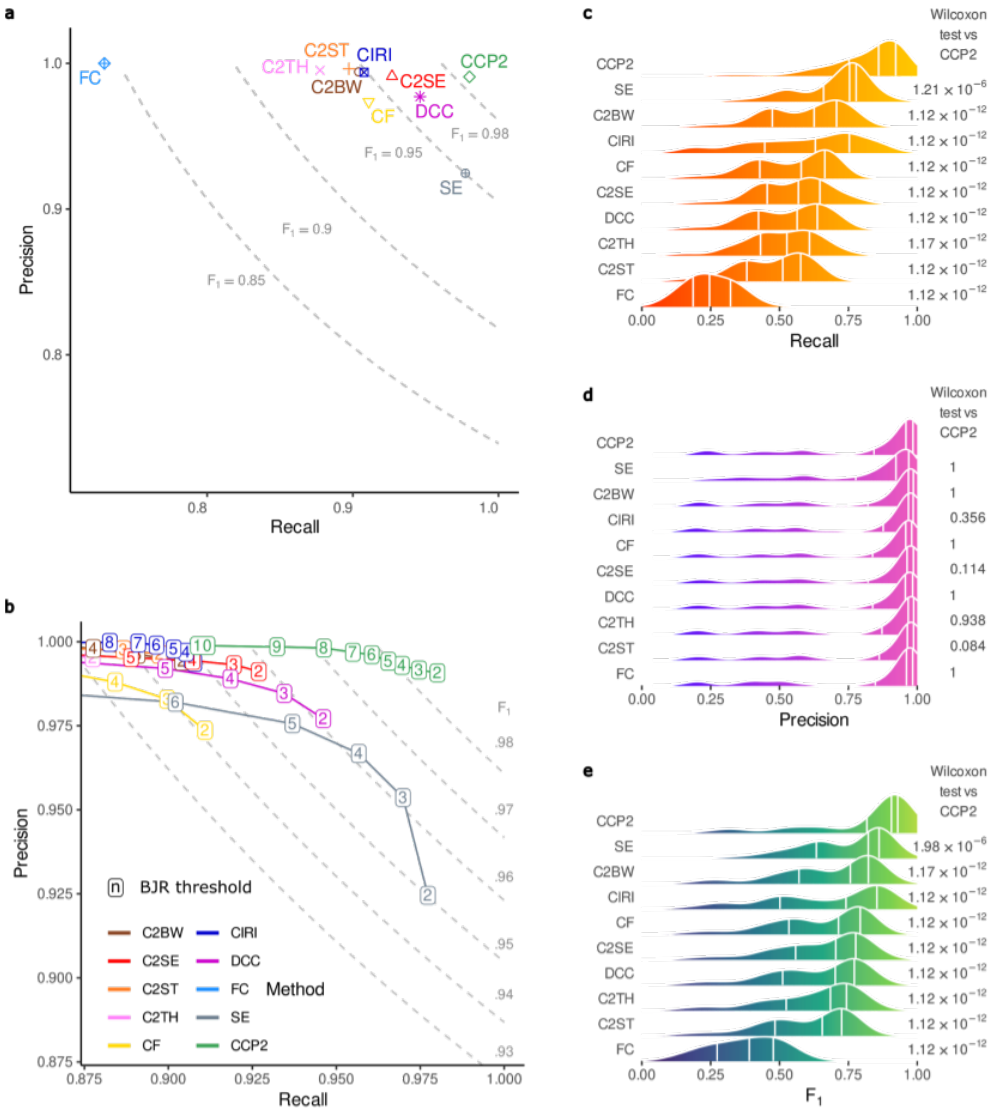

Figure 2. Performance of circRNA detection methods on simulated and real data sets. Simulated data (**a-b**): **a**, precision and recall of the circRNA detection methods (labels as in Figure 1), including CirComPara2 (CCP2; green dot); the dashed-line curves display plot areas for 0.85, 0.90, 0.95 and 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); plot areas of F_1-scores from 0.92 to 0.98 by 0.01 increase step are delimited by dashed-line curves; method colours as in (a). Real data set benchmark density plot of (**c**) recall, (**d**) precision and (**e**) F_1-score; distribution quartiles are indicated by white vertical lines; each method contrasted with CirComPara2 Wilcoxon paired test, Bonferroni adjusted p-values are reported on the right of the plot.

14

References

1. Xiao, M.-S., Ai, Y. & Wilusz, J. E. Biogenesis and Functions of Circular RNAs Come into Focus. *Trends Cell Biol.* **30**, 226–240 (2020).
2. Chen, L. *et al.* The bioinformatics toolbox for circRNA discovery and analysis. *Brief. Bioinform.* (2020) doi:10.1093/bib/bbaa001.
3. Zeng, X., Lin, W., Guo, M. & Zou, Q. A comprehensive overview and evaluation of circular RNA detection tools. *PLoS Comput. Biol.* **13**, e1005420 (2017).
4. Chen, I., Chen, C.-Y. & Chuang, T.-J. Biogenesis, identification, and function of exonic circular RNAs. *Wiley Interdiscip. Rev. RNA* **6**, 563–579 (2015).
5. Hansen, T. B. Improved circRNA Identification by Combining Prediction Algorithms. *Front Cell Dev Biol* **6**, 20 (2018).
6. Hansen, T. B., Venø, M. T., Damgaard, C. K. & Kjems, J. Comparison of circular RNA prediction tools. *Nucleic Acids Res.* **44**, e58 (2016).
7. Westholm, J. O. *et al.* Genome-wide analysis of drosophila circular RNAs reveals their structural and sequence properties and age-dependent neural accumulation. *Cell Rep.* **9**, 1966–1980 (2014).
8. Gao, Y., Zhang, J. & Zhao, F. Circular RNA identification based on multiple seed matching. *Brief. Bioinform.* **19**, 803–810 (2018).
9. Cheng, J., Metge, F. & Dieterich, C. Specific identification and quantification of circular RNAs from sequencing data. *Bioinformatics* **32**, 1094–1096 (2016).
10. Memczak, S. *et al.* Circular RNAs are a large class of animal RNAs with regulatory potency. *Nature* **495**, 333–338 (2013).
11. Hoffmann, S. *et al.* A multi-split mapping algorithm for circular RNA, splicing, trans-splicing and fusion detection. *Genome Biol.* **15**, R34 (2014).
12. Zhang, X.-O. *et al.* Diverse alternative back-splicing and alternative splicing landscape of circular RNAs. *Genome Res.* **26**, 1277–1287 (2016).
13. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. (2013).
14. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).

15

275 15. Kim, D. & Salzberg, S. L. TopHat-Fusion: an algorithm for discovery of novel fusion transcripts.
276 *Genome Biol.* **12**, R72 (2011).
277 16. Gaffo, E., Bonizzato, A., Kronnie, G. T. & Bortoluzzi, S. CirComPara: A Multi-Method
278 Comparative Bioinformatics Pipeline to Detect and Study circRNAs from RNA-seq Data.
279 *Noncoding RNA* **3**, (2017).
280 17. Szabo, L. & Salzman, J. Detecting circular RNAs: bioinformatic and experimental challenges.
281 *Nature Reviews Genetics* vol. 17 679–692 (2016).
282 18. Vromman, M., Vandesompele, J. & Volders, P.-J. Closing the circle: current state and
283 perspectives of circular RNA databases. *Brief. Bioinform.* (2020) doi:10.1093/bib/bbz175.
284 19. Wu, W., Ji, P. & Zhao, F. CircAtlas: an integrated resource of one million highly accurate
285 circular RNAs from 1070 vertebrate transcriptomes. *Genome Biol.* **21**, 101 (2020).
286 20. Ruan, H. *et al.* Comprehensive characterization of circular RNAs in ~ 1000 human cancer cell
287 lines. *Genome Med.* **11**, 55 (2019).