OXFORD

Briefings in Bioinformatics, 2023, **24(1)**, 1–15
https://doi.org/10.1093/bib/bbac612
Problem Solving Protocol

Systematic benchmarking of statistical methods to assess differential expression of circular RNAs

Alessia Buratin, Stefania Bortoluzzi† and Enrico Gaffo†

Corresponding author. Enrico Gaffo, Department of Molecular Medicine, University of Padova - Via G. Colombo, 3—35131 Padova, Italy. Phone +39 049 827 6502; Fax +39 049 827 6209; E-mail: enrico.gaffo@unipd.it

†Stefania Bortoluzzi and Enrico Gaffo contributed equally.

Abstract

Circular RNAs (circRNAs) are covalently closed transcripts involved in critical regulatory axes, cancer pathways and disease mechanisms. CircRNA expression measured with RNA-seq has particular characteristics that might hamper the performance of standard biostatistical differential expression assessment methods (DEMs). We compared 38 DEM pipelines configured to fit circRNA expression data's statistical properties, including bulk RNA-seq, single-cell RNA-seq (scRNA-seq) and metagenomics DEMs. The DEMs performed poorly on data sets of typical size. Widely used DEMs, such as DESeq2, edgeR and Limma-Voom, gave scarce results, unreliable predictions or even contravened the expected behaviour with some parameter configurations. Limma-Voom achieved the most consistent performance throughout different benchmark data sets and, as well as SAMseq, reasonably balanced false discovery rate (FDR) and recall rate. Interestingly, a few scRNA-seq DEMs obtained results comparable with the best-performing bulk RNA-seq tools. Almost all DEMs' performance improved when increasing the number of replicates. CircRNA expression studies require careful design, choice of DEM and DEM configuration. This analysis can guide scientists in selecting the appropriate tools to investigate circRNA differential expression with RNA-seq experiments.

Keywords: CircRNA, differential expression, benchmark, RNA-seq, low-expression

Introduction

Circular RNAs (circRNAs) are transcripts in which an upstream 5′ splice site and a downstream 3′ splice site are covalently joined through a backsplicing process [1]. CircRNAs are pervasively expressed in eukaryotes, play critical cellular roles, are involved in disease and cancer mechanisms, and find many biomedical applications [1–4]. The last decade has seen mounting interest in studying circRNAs [5]. CircRNAs are often investigated through high-throughput total RNA sequencing (RNA-seq), and their characterization is becoming a fundamental part of transcriptomics analyses [6–9].

Most bioinformatics tools that quantify circRNA expression from RNA-seq data estimate circRNA abundance by counting the backspliced reads, i.e. the spliced reads aligned in non-collinear order to backsplice junctions [5, 10]. They allow composing of backsplice junction read (BJR) count expression matrices, which can be analysed with statistical methods devised to assess differential gene expression (DGE) [11]. Although several works benchmarked the numerous DGE assessment tools developed for RNA-seq technology [12–17], circRNA expression data have never been considered so far.

CircRNA expression is generally low, as backsplicing is rarer than linear splicing [1, 18, 19], and, in RNA-seq data, the backspliced reads constitute no more than 2% of all spliced reads [18]. Moreover, technical aspects of the procedure to quantify the circRNA abundance from RNA-seq data may hamper the estimation of circRNA expression levels, even when circRNA-enriched sequencing libraries are employed [19]. Such biological and technical characteristics lead to numerous low-count expression estimates, which can undermine the performance of DGE assessment methods [16, 20, 21].

Thus far, only the circMeta package [22] provides a statistical method specific for differential circRNA abundance. The authors of circMeta observed that the Poisson distribution modelled circRNA expression counts better than the negative binomial (NB). Therefore, they proposed using a Poisson-based z-test to determine differentially expressed circRNAs and showed that it was more powerful than DESeq2 [23] and edgeR [24]. However, their comparison was limited to parametric simulations based on a single real data set with a small number of replicates. Moreover, the parameter settings exploration was limited to the default for the two competitor methods, and only circRNAs expressed at

Alessia Buratin is a PhD student in Biosciences (curriculum Genetics, Genomics and Bioinformatics) of the University of Padova. Her main interests are biostatistics and bioinformatics, transcriptomics of hematologic malignancies, circular RNA function prediction and biogenesis.
Stefania Bortoluzzi is an associate professor of Applied Biology at the Department of Molecular Medicine of the University of Padova, where she leads the Computational Genomics Laboratory. Her research interests include cancer genomics and transcriptomics, bioinformatics, systems biology, noncoding RNAs, circular RNAs, exosomal RNAs and hematologic malignancies.
Enrico Gaffo is a post-doc at the Computational Genomics Laboratory at the Department of Molecular Medicine, University of Padova. His research interests include circular RNA, microRNA, advanced methods for RNA-seq data analysis and bioinformatics applied to cancer research.
Received: September 26, 2022. **Revised:** November 28, 2022. **Accepted:** December 11, 2022
© The Author(s) 2023. Published by Oxford University Press.
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (https://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

Downloaded from https://academic.oup.com/bib/article/24/1/bbac612/6966517 by guest on 19 September 2024

moderate to high levels were considered, as is expected when selecting the circRNAs jointly predicted by multiple circRNA detection tools [25, 26].

In this work, we first explored the characteristics of circRNA expression count data, confirming that most circRNAs yield small counts in typical RNA-seq data sets and highlighting a considerable fraction of zero counts. Then, we compared traditional DGE tools for bulk RNA-seq applied to circRNA expression data, considering different parameter combinations for low-count and sparse data. Because we observed similar features between circRNA, single-cell and microbiome RNA-seq data, we also included statistical models developed for those fields. In total, we compared 38 differential expression assessment pipelines on hundreds of semi-parametric and non-parametric simulated circRNA expression count data sets, evaluating the type I error control, FDR, recall, F_1-score, area under the precision-recall curve (AUPRC) and similarity of predictions between the methods. Our systematic and comprehensive benchmarking provides an overview of the weaknesses and strengths of differential expression tools on circRNA data.

Results

CircRNA expression data are characterized by a high proportion of small and zero counts

The signal to measure circRNA abundance from RNA-seq is limited compared to that available for linear transcripts and genes. CircRNAs are generally less abundant than linear transcripts and can be quantified unambiguously only by the reads encompassing the backsplice junctions [10, 27] (Figure 1A). Besides, gene expression abundance is measured by counting both the spliced and unspliced reads aligned to the whole gene region, thus summing the expression of all transcript isoforms of a gene [28]. In contrast, each circRNA represents one single transcript, and the BJRs originate only from the specific site of the circRNA sequence where the junction ends were joint [19] (Figure 1A). Moreover, BJRs are computationally harder to identify than unspliced and linearly spliced reads as they require non-collinear alignments and additional processing to remove spurious hits [5], causing most circRNA detection tools to suffer from low detection rates [19, 26].

The combination of circRNA biological features and computational hindrances in estimating their expression can result in data sets with a large fraction of small counts. We verified this characteristic in 34 RNA-seq data sets of matched ribosomal RNA-depleted and circRNA-enriched libraries from 17 human tissues (Table 1).

CirComPara2 [26] was used to obtain linear and circular read mappings on circRNA-host genes. We discriminated four read alignment sets representing the expression signal available for (i) estimating gene expression, (ii) studying alternative splicing, (iii) comparing the abundance of circular and linear transcripts expressed by a gene, and (iv) estimating circRNA abundance. We compared the magnitude of the expression signals by counting (i) the unspliced and linearly spliced reads together, (ii) only the linearly spliced reads, (iii) only the linearly spliced reads aligned into backsplice junctions and (iv) only the BJRs (Figure 1A).

Regardless of the circRNA library enrichment, we observed that the highest signal was obtained for gene expression estimates, followed by the linearly spliced reads (Figure 1B; Supplementary Figure S1). In turn, the spliced read counts slightly diminished if considering only those mapped on backsplice junctions. The BJRs showed the lowest values, also in the circRNA-enriched samples (Figure 1B). Notably, median BJR counts were less than or equal to 10 in most samples (Supplementary Figure S1). These observations supported the hypothesis that, in RNA-seq data, circRNA expression estimates rely on a low signal biased by the quantification procedure.

We further considered circRNA expression in RNA-seq data sets with multiple biological replicates to analyse the BJR count distribution of circRNA expression matrices. From the Sequence Read Archive public repository [29], we collected RNA-seq data of four independent circRNA studies that compared groups with at least five samples sequenced, more than 40 million paired-end reads, and composed of 10–50 samples of human tumours and healthy tissues (Table 1). In each data set, the samples showed high circRNA expression correlation within conditions, denoting homogeneity of the samples (Table 1).

In these data sets, most BJR counts laid below 10 (Figure 1C), indicating that the circRNA small counts were not data set specific. Moreover, most circRNAs had a median BJR count of less than 10 (Figure 1D), and the more samples in which a circRNA was detected, the higher the median BJR count (Figure 1E).

These results suggested that the less expressed circRNAs might be undetected in some samples because of a sampling bias [30], which could inflate zero counts. Notably, the zero-count fraction was large in all data sets (Figure 1C). The unfiltered data set sparsity ranged from 44 to 72% null counts, but was not significantly correlated to the library size ($r_{\text{Pearson}} = 0.05$, P-value >0.9) (Supplementary Figure S2).

To ascertain that these observations were not determined by some artefacts of the circRNA expression estimation algorithm, we computed the BJR counts with six additional circRNA quantification pipelines. We observed that the BJR count distribution was comparable among the quantification methods (Supplementary Figure S3), and the proportion of zero counts was high in all data sets regardless of the quantification pipeline (Figure 1F).

Plus, as it is common practice in RNA-seq expression analysis [31], we applied five independent expression filtering strategies to the BJR count matrices, which considered discarding circRNAs according to the number of samples in which they were detected. As expected, the expression filters reduced the number of zero counts, but at the cost of discarding a significant fraction (from 30% up to 95%) of circRNAs (Figure 1F). Moreover, the number of the low BJR counts remained high (Supplementary Figure S3), suggesting that the circRNAs detected in multiple samples also yielded a low expression signal.

Statistical modelling of circRNA expression count data

RNA-seq count data is often modelled with an NB distribution [32]. However, when zero counts are in excess, a zero-inflated NB distribution (ZINB) may fit the data better [33]. We thus evaluated whether a ZINB distribution can model BJR counts better than an NB by calculating the goodness-of-fit (GOF) on the BJR count matrices, unfiltered and upon applying the expression filters. For each circRNA, the NB and ZINB GOF were compared according to the root mean square error (RMSE) of the mean counts and probability of observing a zero and the Akaike information criterion (AIC) scores.

Both NB and ZINB distributions obtained a small error for the mean count estimation (RMSE <0.07) independently of the expression filtering procedure and data set (Supplementary Figure S4). The ZINB model provided better estimates of the observed zero proportion than the NB for each expression filter and data set

Downloaded from https://academic.oup.com/bib/article/24/1/bbac612/6966517 by guest on 19 September 2024

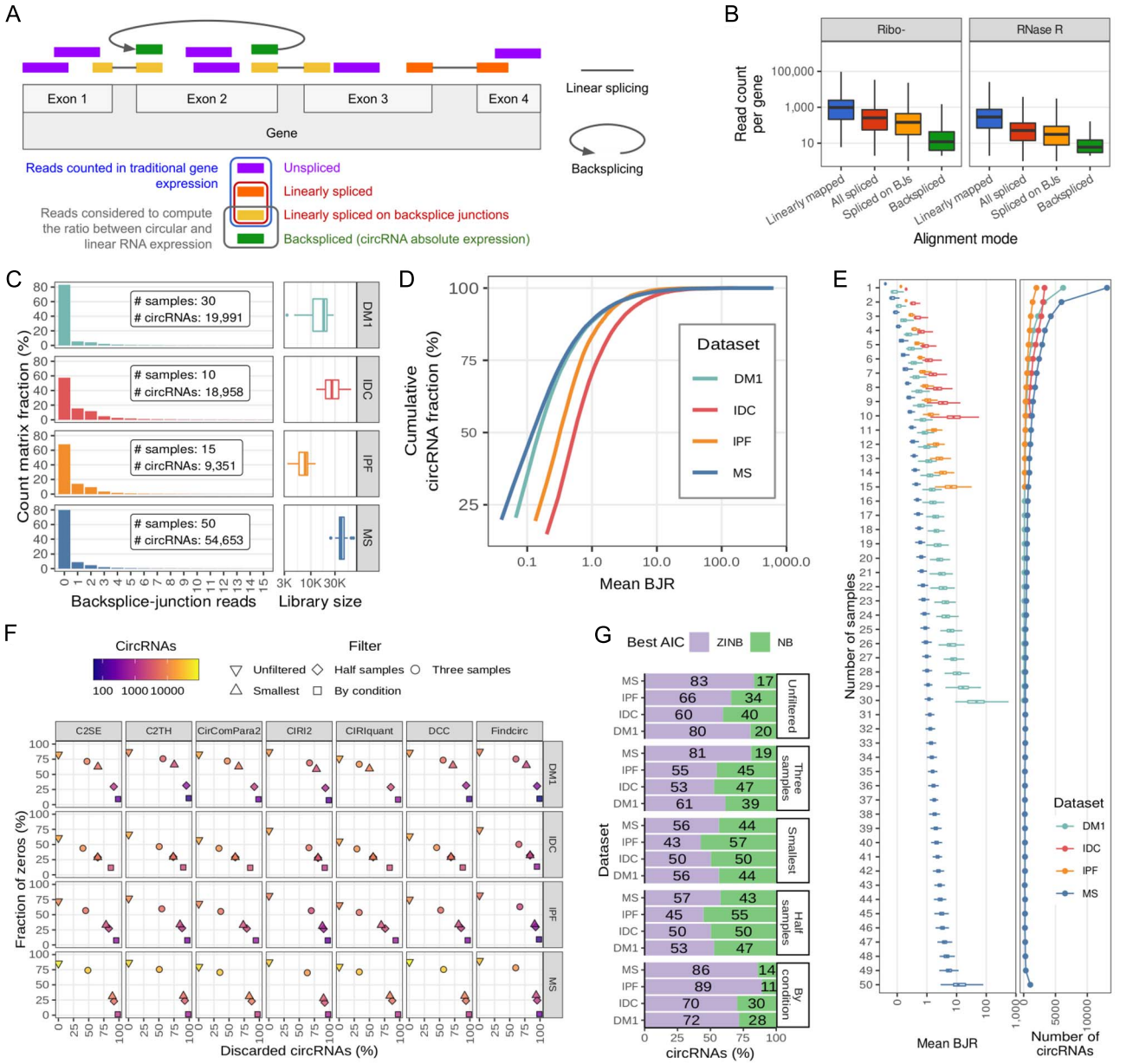

Figure 1. BJR count characteristics and circRNA expression in data sets. (**A**) The reads used to estimate linear transcript and gene expression in conventional RNA-seq gene expression analysis (blue frame) include unspliced reads (purple) and linearly spliced reads (red frame). The reads available to analyse the circRNA expression (grey frame) include the backspliced reads (green), which are used to estimate circRNA absolute expression, and the subset of the linearly spliced reads that involve backsplice junctions (yellow), which are compared to the backspliced reads to compute relative abundance of a circRNA to the host-gene linear expression. (**B**) The distribution of the per-gene mapped read counts in human brain tissue samples (PRJCA000751) sequenced from rRNA-depleted RNA-seq libraries (Ribo-) and Ribo- followed by RNase R treatment (RNase R). Four sets of reads, corresponding to the groups and colours as in (**A**), were considered: Linearly mapped (blue), All spliced (red), Spliced on BJs (yellow) and Backspliced (green). (**C**) The proportion of BJR counts in the multi-sample DM1, IDC, IPF and MS data sets. The number of samples composing each data set and the total number of circRNAs detected is reported. The boxplot of each data set's sample library size is on the right. (**D**) The cumulative fraction of circRNAs for expression levels measured as average BJR. (**E**) The circRNA expression distribution in average BJR, given the number of samples in which the circRNAs were detected. (**F**) Percentage of zero counts and the number of circRNAs discarded upon the filtering procedure for each detection tool and data set. (**G**) The percentage of circRNAs best modelled by an NB or a ZINB model according to AIC in each data set and upon the filtering procedures.

except in the MS data set upon the application of the two filters discarding most circRNAs (i.e. 'By condition' and 'Half samples') (Supplementary Figure S5). However, we observed small errors also for the NB (RMSE <0.09). According to the AIC measure, the ZINB distribution modelled the BJR count data better than the NB for most circRNAs across data sets and expression filters (mean 68 ± 17%) (Figure 1G), suggesting that the model accounting for an excess of zeros might improve fitting circRNA expression data.

Comparison of differential expression assessment methods on circRNA data

In this work, we focus on the problem of assessing circRNA differential abundance. The traditional methods for bulk RNA-seq data analysis have been the primary choice when analysing circRNA expression, with DESeq2 [23], edgeR [24, 31, 34] and Limma-Voom [35] arguably the most used. However, the circRNA expression characteristics shown above suggest that circRNA BJR count data

Downloaded from https://academic.oup.com/bib/article/24/1/bbac612/6966517 by guest on 19 September 2024

Table 1. The characteristics of the circular RNA data sets analysed in this study

Name in this study	Accession ID and original study reference	Use in this study	Number of samples; replicates in each group; sample tissue; min-max sequenced reads; read type	Correlation among replicates*
JHS	[a]PRJCA000751 [29]	Comparison of read counts per alignment type	17 human tissue samples, matched ribodepleted and RNase-R treated libraries; 31–254 M 150 bp PE	-
DM1	§§GSE86356 [30]	Semi-parametric simulations	30 samples: 25 Myotonic Dystrophy Type 1 and 5 tibialis anterior muscle biopsies (healthy controls); 84–120 M 50 bp PE	(0.76, 0.96)
IDC	§§§SRP156355 [31]	Semi-parametric simulations	10 samples: 5 Invasive Ductal Carcinoma and 5 normal breast tissue; 67–120 M 100 bp PE	(0.42, 0.89)
IPF	§§GSE52463 [32]	Semi-parametric simulations	15 samples: 8 Idiopathic pulmonary fibrosis and 7 normal; 40–60 M 100 bp PE	(0.50, 0.94)
MS	§§GSE159225 [33]	Semi-parametric simulations	50 samples: 30 multiple sclerosis and 20 healthy tissues; 80–140 M PE 150 bp	(0.74, 0.77)
PC	PMID 35078526 [34]	Non-parametric simulations	96 samples: 20 normal versus 76 prostate cancer (only batch 2 of the original data set); CIRI2 BJR count matrix	(0.40, 0.90)

[a]NGDC ID (National Genomics Data Center, China National Center for Bioinformation); §§GEO ID; §§§SRA ID; * Pearson's correlation (minimum, maximum) calculated among replicates within conditions based on BJR counts; M: million reads; PE: paired-end; BJR: backsplice junction read

could not comply with the traditional differential expression methods (DEMs) assumptions and disrupt their performance.

The high proportion of small counts and the sparsity of circRNA expression data are comparable to those observed in single-cell RNA-seq (scRNA-seq) and whole metagenome shotgun sequencing (WMS). In particular, the circRNA data's small counts and library size are similar to droplet-based scRNA-seq data [30]. Further, the sparsity of circRNA data is analogous to scRNA-seq and WMS data, which range between 12 and 75% zeroes and 35 and 89%, respectively [17]. Finally, we observed that a ZINB distribution fits circRNA data better than NB in half the cases, as described in full-length scRNA-seq [30].

Therefore, we benchmarked 18 DEMs, including bulk RNA-seq DEMs and a few tools conceived for scRNA-seq and WMS data, selecting those freely available as R packages or functions. Furthermore, we explored different parameter settings, the ZINB-WaVE package weighting strategy [36, 37] and normalization approaches [38, 39] coupled to DESeq2, edgeR and Limma-Voom to handle small counts and sparse data specifically.

In total, we compared 38 differential expression analysis pipelines (Supplementary Table S1), evaluating their type I error control, FDR, true positive rate (TPR, or recall), F_1-scores, AUPRC and computation time. Moreover, we calculated the similarity of predictions between DEMs according to two similarity indexes.

Benchmark data sets simulation with a semi-parametric approach

We generated 720 simulation data sets using SP-SimSeq [40], a semi-parametric approach that preserves the real circRNAs and circRNA-circRNA correlations observed in real data. Specifically, for each of the four multiple-sample data sets, we simulated 30 expression matrices with an equal number of samples in two conditions considering three (N03), five (N05) and ten (N10) samples per group. 'Null' data sets with no differentially expressed circRNAs (DECs) and 'signal' data sets with 10% DECs were generated. We evaluated the simulated data sets' quality according to expression levels, fractions of zeros and the relation between the both, as in Soneson and Robinson [41]. All measures were not significantly different from the original data sets, confirming that the simulated data followed the original data characteristics (Supplementary Tables S2–S4).

The following paragraphs show the results of the N05 size data sets for the 0.05 significance threshold. We reasoned that this is a common scenario for circRNA RNA-seq experiments. The results from the N03 and N10 simulations at 0.01 or 0.1 significance are available in the Supplementary Material.

Type I error control

We evaluated the type I error rate for each DEM, i.e. the probability of predicting a DEC when it is not, by computing the false positive rate (FPR) in the 'null' data sets. The 'null' data sets allowed to evaluate the DEMs' type I error rate as any DEC called as significant represent a false-positive (FP) prediction. The methods could be grouped according to (i) liberal, (ii) conservative and (iii) sufficient control of the type I error (Figure 2). Among the liberal methods, Seurat-BIM-LRT showed largely uncontrolled type I error (FPR = 0.38), consistently with a previous assessment by Soneson and Robinson in scRNA-seq [15]. Other methods with moderately liberal type I error control included Seurat-WLX and three edgeR pipelines (TWSP, RBST and 50DF), with a median FPR between 0.10 and 0.12, whereas slightly liberal methods included edgeR-ZW, NBID, Voom-LF, Voom-QN ($0.07 \leq FPR \leq 0.08$). In contrast, conservative results ($0 \leq FPR < 0.03$) were obtained by MAST, lncDIFF, DEsingle, the Wilcoxon test, Limma-VST and all the DESeq2 pipelines but DESeq2-ZW. The remaining methods achieved an FPR close to the nominal value ($0.03 \leq FPR < 0.07$), and most of them (10 out of 16) were bulk RNA-seq methods. Results from NOISeqBIO were not suitable for type I error estimation because the NOISeqBIO's scores are comparable with adjusted *P*-values, explaining its low FPR. We show NOISeqBIO in this analysis only for the completeness of the report. The quasi-likelihood framework [31, 42], devised to improve type I error control when a linear model contains fitted values that are exactly zero, was effective using edgeR (edgeR-QFT) but not with the Limma-Voom pipeline (Voom-LF), which obtained slightly higher FPR than the other Limma-Voom versions. Interestingly, DESeq2 showed a type I error closer to the imposed α only when using the ZINB-WaVE weights.

The performance of the methods was consistent regardless of the α threshold (Supplementary Figure 6). Plus, a larger sample set improved the error rates only slightly for most methods except DESeq2, especially DESeq2-ZW, which showed much better results in larger data sets. Instead, Seurat-WLX, metagenomeSeq and the three edgeR pipelines mentioned above showed larger FPRs with data sets of increased size (Supplementary Figure S6).

Downloaded from https://academic.oup.com/bib/article/24/1/bbac612/6966517 by guest on 19 September 2024

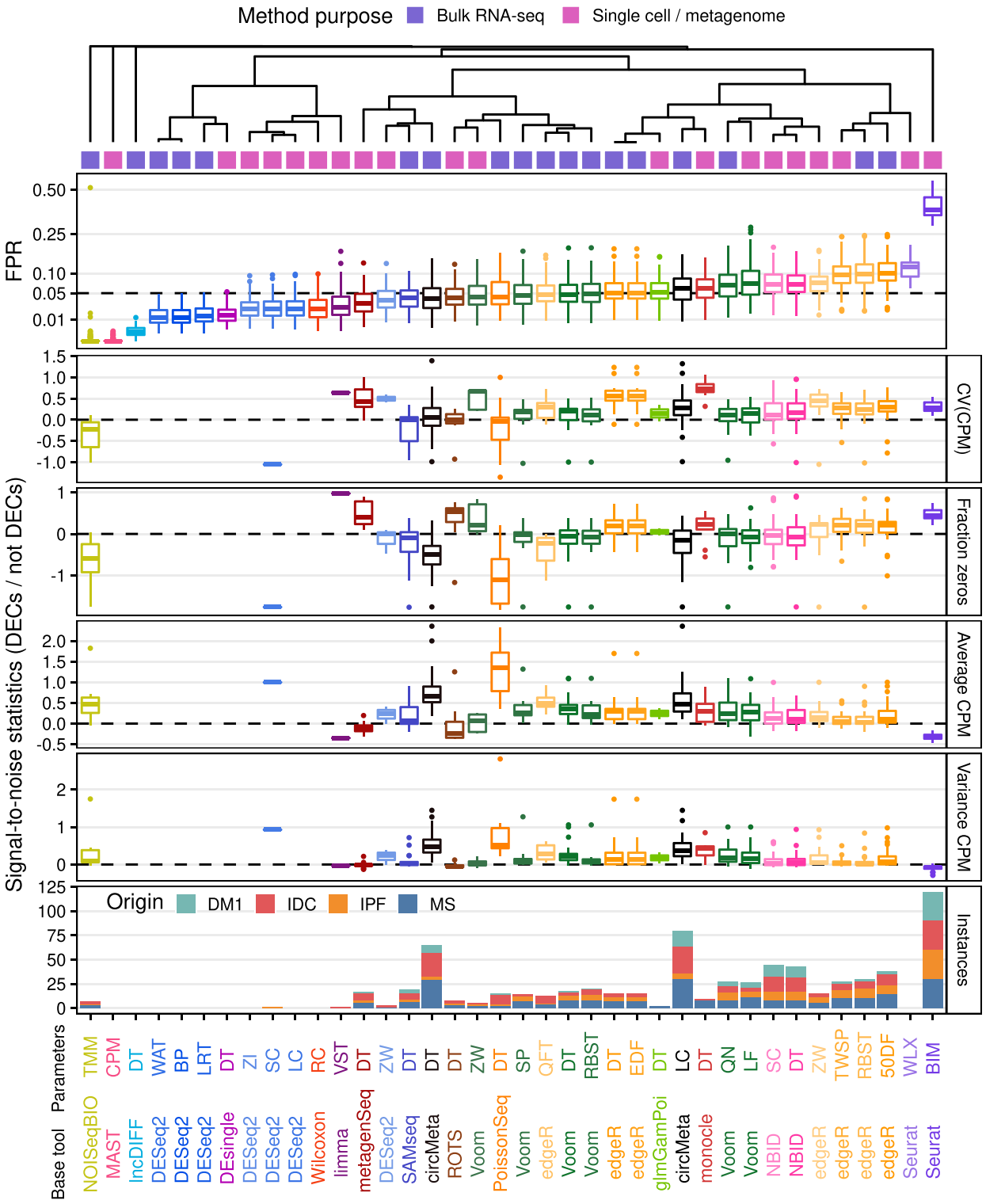

Figure 2. Type I error control and signal-to-noise statistic rates. FPR at *P*-value 0.05 of the DEMs on the 'null' data sets with 10 samples (N05). On top, the clustering according to FPR difference from the 0.05 nominal alpha using Euclidean distance and complete agglomeration method; indication of the method purpose; and boxplots of the FPR scores. The central boxplots show the signal-to-noise statistics (coefficient of variation, CV; fraction of zero counts; average and variance of counts per million; CPM) of circRNAs called differentially expressed (DECs) at adjusted *P*-value 0.05. The bottom stacked bar chart shows the number of simulations with at least five DECs, colouring according to the data set from which the instances were simulated. Bulk: tools devised for bulk RNA-seq data; Single-cell/metagenome: tools devised or adapted for single cell and metagenome RNA-seq data.

Downloaded from https://academic.oup.com/bib/article/24/1/bbac612/6966517 by guest on 19 September 2024

Expression estimate characteristics of the FP differentially expressed circRNAs

We calculated signal-to-noise statistics for each tool that reported five or more FP DECs in at least one 'null' data set. Similarly to a previous work by Soneson and Robinson [15], we compared the significant and non-significant DECs according to their average counts per million (CPM), coefficient of variation (CV), variance and mean fraction of zeros (Figure 2). In general, we did not observe marked signal-to-noise statistics. The CV and variance of CPM were slightly positive for all methods, particularly the edgeR pipelines, except the CV for SAMseq and PoissonSeq, which were mostly negative. Likewise, the average CPM of FP DECs was slightly higher than not significant circRNAs, except for Seurat-BIM, ROTS, metagenomeSeq and the few FP DECs predicted by Limma-VST. We observed a heterogeneous behaviour regarding the fraction of zeros: Limma-VST, metagenomeSeq, ROTS, Voom-ZW, monocle, edgeR (except edgeR-QFT) and Seurat-BIM failed more on circRNAs with higher fractions of zero counts. FP DECs originated approximately equally from all scenarios except for circMeta-DT, which showed a higher FP number on the IDC and MS data set instances (Figure 2). Overall, the zero counts did not significantly affect the type I error control as much as the expression abundance and variance.

FDR, power, F_1-score and AUPRC

We used the 'signal' data sets to evaluate the methods' FDR and TPR. The Wilcoxon's test and Seurat-WLX did not generate any significant prediction (Figure 3A), thus resulting in a null TPR (Figure 3B), and lncDIFF, MAST and DEsingle returned significant predictions only in a few simulation instances. Similarly, DESeq2, PoissonSeq and Limma-VST did not predict any DEC in a relevant fraction of simulation instances, especially from the MS data sets. Only Seurat-BIM and circMeta-LC provided predictions below the imposed level in all the simulation instances.

Most methods (22 out of 38) showed higher FDR than the imposed 0.05 level: lncDIFF and Seurat-BIM scored the worst FDRs (FDR = 1 and 0.95, respectively), followed by PoissonSeq, circMeta, all edgeR pipelines, NBID, metagenomeSeq, monocle, ROTS and all Voom pipelines but Voom-ZW (FDR > 0.09). In contrast, NOISeqBIO, MAST, DEsingle, Limma-VST and all DESeq2 pipelines but DESeq2-ZW controlled the FDR lower than the nominal value ($0 \leq \text{FDR} \leq 0.01$). In DESeq2, a slightly more conservative FDR was obtained using the likelihood ratio test (DESeq2-LRT) compared with the Wald test (DESeq2-WaT). Voom-ZW, DESeq2-ZW, glmGamPoi and SAMseq controlled the FDR close to the nominal value. Notably, every method except DEsingle and MAST presented FDP close to 1 in some instances. All methods obtained better FDR control on the N10 data sets but maintained the characteristics observed in the N05 data sets (Supplementary Figure S7).

The sensitivity was generally low, with a median below 50% for all methods (Figure 3B). The highest TPRs ($0.43 \leq \text{TPR} \leq 0.41$) were obtained by three edgeR pipelines (TWSP, RBST and 50DF). SAMseq, NBID, monocle and four Voom pipelines (Voom-QN, Voom-LF, Voom-RBST and Voom-DT) obtained TPRs between 0.36 and 0.31, whereas the remaining methods identified less than 30% true DECs. The choice of parameters greatly influenced sensitivity in edgeR pipelines. Interestingly, the quasi-likelihood framework produced opposite results when applied to edgeR or Limma-Voom, with the lowest and the highest TPR among the respective pipeline configurations. Similarly, ZINB-WaVE weights allowed higher recall rates with edgeR and DESeq2 but a lower TPR with Limma-Voom. Regarding the DESeq2 pipelines, the scRNA-seq-oriented pipelines obtained higher recall rates than the bulk RNA-seq configurations (Figure 3B; Supplementary Figures S8–9). Notably, the adaptation to low counts of the circMeta test sensibly improved the recall rate (median TPR 0.23 and 0.03, respectively). Poor performance, close to zero, was achieved by Limma-VST, lncDIFF, DEsingle, MAST, the Wilcoxon test, Seurat-WLX and the bulk RNA-seq configurations of DESeq2.

All methods achieved significantly higher sensitivity with increased set sizes, except lncDIFF and Seurat-WLX, which did not detect any true DEC, and metagenomeSeq, which improved only a little (Supplementary Figure S9). The highest TPR among all settings (TPR = 0.9) was achieved by edgeR-RBST and edgeR-TWSP when allowing for a 0.1 adjusted *P*-value threshold in the N10 data sets (Supplementary Table S5). In the smallest data sets (N03), NOISeqBIO had the highest recall rate (TPR = 0.7 with 0.1 adjusted *P*-value), which was surprisingly higher than in the larger sets (Supplementary Figure S9).

We inspected the *P*-value distribution obtained in the 'signal' data sets to understand better the DEMs' predictions (Supplementary Figure S10). CircMeta, edgeR, glmGamPoi, Limma-VST, NBID, PoissonSeq, ROTS, SAMseq and Voom showed *P*-value histograms as expected [43]. The other DEMs did not show a uniform distribution of the *P*-values, most having an overabundance of large *P*-values or a distribution skewed towards $P = 1$. Interestingly, the DESeq2 overabundance of large *P*-values was mitigated using the weights for zero counts. Comparing the *P*-value histograms between the N05 (Supplementary Figure S10) and N10 (Supplementary Figure S11) simulations, we observed better *P*-value distributions, indicating that the conservative *P*-value distributions were due to insufficient power of the methods with a small number of samples [16, 43]. We observed a worse performance of Seurat-WLX compared to the simple Wilcoxon rank-sum test. As Seurat-WLX implements an extended Wilcoxon rank-sum test that considers correlations between cases, the presence of positive correlations between circRNAs possibly increased the variance of the test, making the test more conservative.

Similarly to the analysis of type I error, we calculated the signal-to-noise statistics of the variability, fraction of zeros and expression abundance, comparing for each method the false-negative (FN) and true-positive (TP) predictions, i.e. the circRNAs not detected as differentially expressed compared to those correctly identified. We did not observe significantly different characteristics of the FN compared to the TP predictions (Supplementary Figures S12–14). The poor recall rate could be related to an imprecise dispersion estimation of the models [44] or a systematic deviation from the theoretical null distribution of the test statistics [43].

We calculated the F_1-score of each method to evaluate precision and recall simultaneously (Figure 3C; Supplementary Table S5). Monocle and SAMseq obtained the highest F_1-score ($F_1 = 0.61$), followed by Voom-QN ($F_1 = 0.58$), edgeR-TWSP and edgeR-RBST ($F_1 = 0.57$ and 0.56, respectively).

We observed that the methods generally achieved better precision than recall and that precision scores were less spread than recall scores. In particular, edgeR-TWSP and edgeR-RBST owed their high F_1-scores mainly to their high recall rates. Instead, SAMseq, monocle and Voom-QN scores were driven mainly by a high precision ($\text{PPV} \geq 0.88$) (Supplementary Table S5). Interestingly, the circMeta tests, designed explicitly for circRNA expression, ranked amongst the lowest F_1-scores. SAMseq held the highest score also in the N03 data sets (Supplementary Figures S15–16). However, we observed a different ranking in the N10 data sets: Voom and

Downloaded from https://academic.oup.com/bib/article/24/1/bbac612/6966517 by guest on 19 September 2024

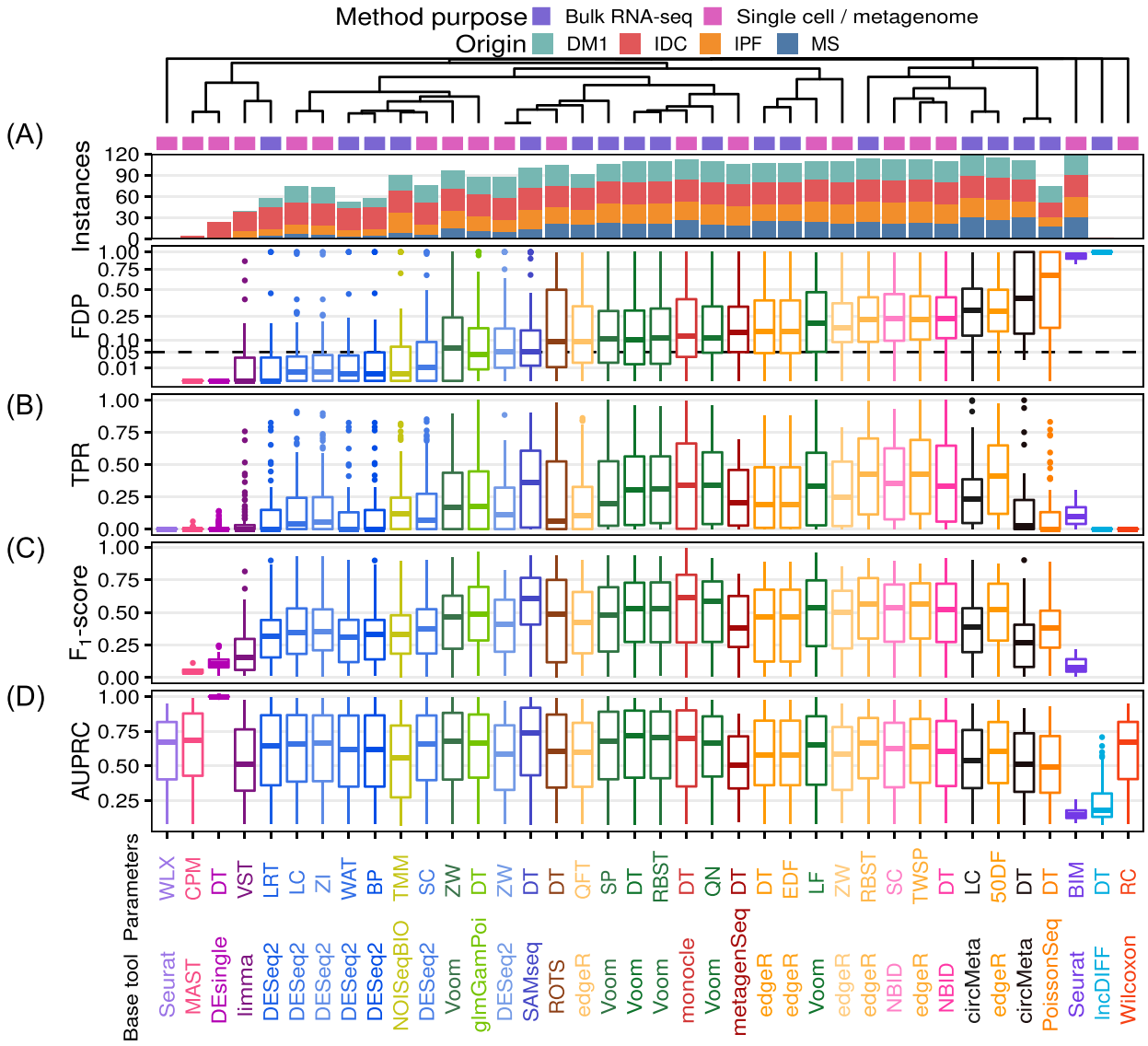

Figure 3. Performance on N05 semi-parametric simulated data sets at adjusted *P*-value 0.05. (**A**) On the top is a dendrogram showing the clustering of the methods according to false discovery proportion (FDP) difference to the nominal value (Canberra distance, optimal leaf ordering), and whether the method was developed for bulk (purple) or single-cell/metagenome (magenta) RNA-seq; the bars below show the number of simulated instances with significant predictions for each data set. The boxplots show the distribution of each method's FDP, (**B**) TPR, (**C**) F_1 score and (**D**) AUPRC obtained in all the simulated data sets. To increase visibility, the FDP y-axis was square-root-scaled, and the AUPRC was exponentially scaled.

monocle still achieved the top scores, but ROTS, glmGamPoi, DEsingle and five edgeR configurations ranked ahead of SAMseq (Supplementary Table S5; Supplementary Figures S15–S16).

Finally, we inspected the ability of the methods to rank true DECs ahead of not significant ones by computing the AUPRC. The AUPRC is informative for data sets with a significant skew in the class distribution [45, 46], as in our simulations. DEsingle obtained the highest AUPRC scores, notwithstanding its poor performance observed in the above analysis, indicating that DEsingle could almost perfectly rank true DECs on the top positions and suggesting an overly conservative assignment of the *P*-values (Figure 3D). SAMseq, Voom-DT, Voom-RBST and monocle obtained the next best scores (median AUPRC ≥0.7) (Supplementary Table S6). Interestingly, we observed that some methods showing poor performance according to the above metrics, including DESeq2-ZI, the Wilcoxon-based methods and MAST, obtained AUPRC scores comparable to the best- performing tools. In larger data sets, only lncDIFF and Seurat-BIM showed small AUPRC scores and a modest improvement (Supplementary Figure S17).

Analysis with non-parametric simulations from an independent data set

To corroborate the outcomes of the semi-parametric simulation analysis, we performed non-parametric simulations from an independent study data set. We obtained the BJRs of 8239 circRNAs computed with CIRI2 in 20 normal tissues and 76 tumour samples from a recent study on prostate cancer [47] (Table 1). In this data set, the number of replicates was sufficient to use the SimSeq tool [48], which performs non-parametric simulations without imposing any distribution assumption on the simulated data. Similarly

Downloaded from https://academic.oup.com/bib/article/24/1/bbac612/6966517 by guest on 19 September 2024

to the previous analysis, we generated 90 instances for 'null' and 'signal' data sets with 6, 10 and 20 samples of two equally large condition groups. We analysed these data sets like the semi-parametric data and ranked the methods according to FPR, TPR and FDR in each simulation type (Supplementary Figures S18–20). We observed a significant positive correlation between the mean ranks of the non-parametric and semi-parametric simulations (Spearman's rho >0.5, *P*-value <0.001; Supplementary Table S7), indicating a generally consistent performance of the methods in the two simulation settings. Unexpectedly, DEsingle showed an opposite AUPRC score than the semi-parametric results.

Similarity of DEMs' predictions

The methods' similarity was explored in the semi-parametric simulations according to two metrics that considered the magnitude of prediction overlap and inquiring into different aspects of the use of predictions. First, we evaluated the similarity between method pairs according to the overlap of their DECs with an adjusted *P*-value ≤0.05, which allowed us to calculate the Jaccard similarity coefficient. Second, for each method pair, we considered the area under the concordance at the top (CAT), which we defined as the overlap of the top 100 circRNAs ranked according to adjusted *P*-values, regardless of fixed thresholds for the adjusted *P*-values.

Clustering the DEMs according to the similarity indexes, we observed that DEMs of the same base tool tended to cluster together (Figure 4). In particular, DESeq2 and Voom showed a high degree of similarity within the respective pipelines, suggesting that modifying the parameters of these tools did not affect their outcomes much. Instead, the three edgeR pipelines characterized by high FPR and TPR clustered apart from the other edgeR configurations. Consistently with the results above, Voom-LF clustered apart from edgeR-QFT. DESeq2 and edgeR using ZINB-WaVe weights reported similar predictions but slightly different from Voom-ZW. Interestingly, edgeR pipelines clustered closer to the Voom than DESeq2 pipelines according to Jaccard similarity (Figure 4A), whereas three edgeR configurations grouped closer to DESeq2 when considering CAT (Figure 4B), indicating more conservative *P*-values provided by DESeq2. Further, the scRNA-seq and bulk RNA-seq DEMs did not show distinct groups, indicating that they can provide similar results.

In the N10 data sets, allowing adjusted *P*-values ≤0.1, the DESeq2 pipelines showed the most consistent predictions regardless of the parameter configuration according to both the Jaccard index and CAT (Supplementary Figures S21–22). Conversely, the other DEMs showed a consistent ranking of their predictions (Supplementary Figures S22) but a great variation according to Jaccard similarity (Supplementary Figures S21), suggesting that the parameter configurations influenced the *P*-value magnitude but maintained the DEC ranking.

Overall ranking of the methods

To compare the methods' performances overall, we computed each method's rank relative to the other DEMs according to the F_1 score, FDR, TPR, AUPRC and FPR measures, independently in each simulated data set, with lower ranks corresponding to better-performing methods. The mean ranks and standard deviations computed on the N05 data sets are represented in Figure 5.

LncDIFF, MAST, Seurat-WLX, the simple Wilcoxon test and DEsingle consistently performed worse than the other methods in all simulations. DEsingle achieved a good ranking according to the AUPRC, but the above analysis showed its unreliable behaviour in different data sets. NOISeqBIO, the DESeq2-BP, DEseq2-LRT, DEseq2-WAT, Limma-VST and PoissonSeq showed poor performance, ranking close to or higher than the third quartile. Seurat-BIM ranked the worst according to AUPRC and FPR. DESeq2 obtained poor ranking according to F_1 score, FDR and TPR while scoring average ranks for the AUPRC and FPR. Interestingly, DESeq2-ZW showed a slightly better ranking than the other DESeq2 configurations.

EdgeR-RBST, edgeR-TWSP, edgeR-50DF and NBID obtained the best mean ranks (below or close to the first quartile) according to F_1 scores, owing mainly to their high TPRs. However, the edgeR pipelines ranked poorly according to FPRs, putting some concerns about the reliability of their predictions. Besides, NBID was outperformed by more than half the DEMs, according to the AUPRC, suggesting that it is suboptimal for modulating a significance threshold. All the Voom pipelines except Voom-ZW obtained rank below the median in all measures, indicating the consistently good performance of the Limma-Voom models, especially Voom-DT and Voom-RBST. The other edgeR-based methods were a close second. SAMseq and monocle showed interesting results on average but with a large variation, which indicates less consistent performance. Notably, all DEMs' mean FPR ranks were above the first quartile, indicating that no method consistently outperformed the others in controlling type I error.

Different rankings were obtained on the data sets with 3 and 10 replicates per group (Supplementary Figure S23), confirming that the sample size greatly influenced the method performances. DESeq2 obtained the best improvement in larger data sets, whereas circMeta, NOISeqBIO and NBID showed better rankings with small numbers of samples.

Computational time

We compared the methods according to the CPU time required for the analysis (Supplementary Figure S24). Most methods ran rapidly in a few seconds or less than one minute. Conversely, computing weights with ZINB-WaVE was the most time-demanding task. Monocle, DEsingle, NBID and ROTS were the slowest methods, requiring 2–8 minutes to complete the analysis of one simulated data set.

Discussion

In this work, we observed that the biological characteristics of circRNA expression and the technical aspects of its abundance estimation from bulk RNA-seq could generate data with a high proportion of 'very small' counts, i.e. with a mean count in the 2–10 range [49]. Moreover, we found a substantial proportion of zero counts in circRNA BJR count matrices, comparable in magnitude with scRNA-seq and metagenome data [17]. These particular circRNA expression properties can violate the assumptions of the traditional DEMs, including DESeq2 and edgeR [50], which are currently used to evaluate also differential circRNA expression.

A high sequencing depth can mitigate data sparsity and proportion of small counts and better quantify circRNA expression levels [27]. However, the more reads are sequenced, the more expensive the experiment is, and still, the inefficient process of BJR estimation hinders detecting and quantifying circRNAs entirely. This study found that DEMs' power improved in larger data sets, particularly with 10 samples per condition. In line with our observations, also previous works showed that DEMs have significantly higher detection power of lowly expressed genes with an increased number of replicates than with an increased sequencing depth [13], and 10 or more replicates per group are advised [51].

Downloaded from https://academic.oup.com/bib/article/24/1/bbac612/6966517 by guest on 19 September 2024

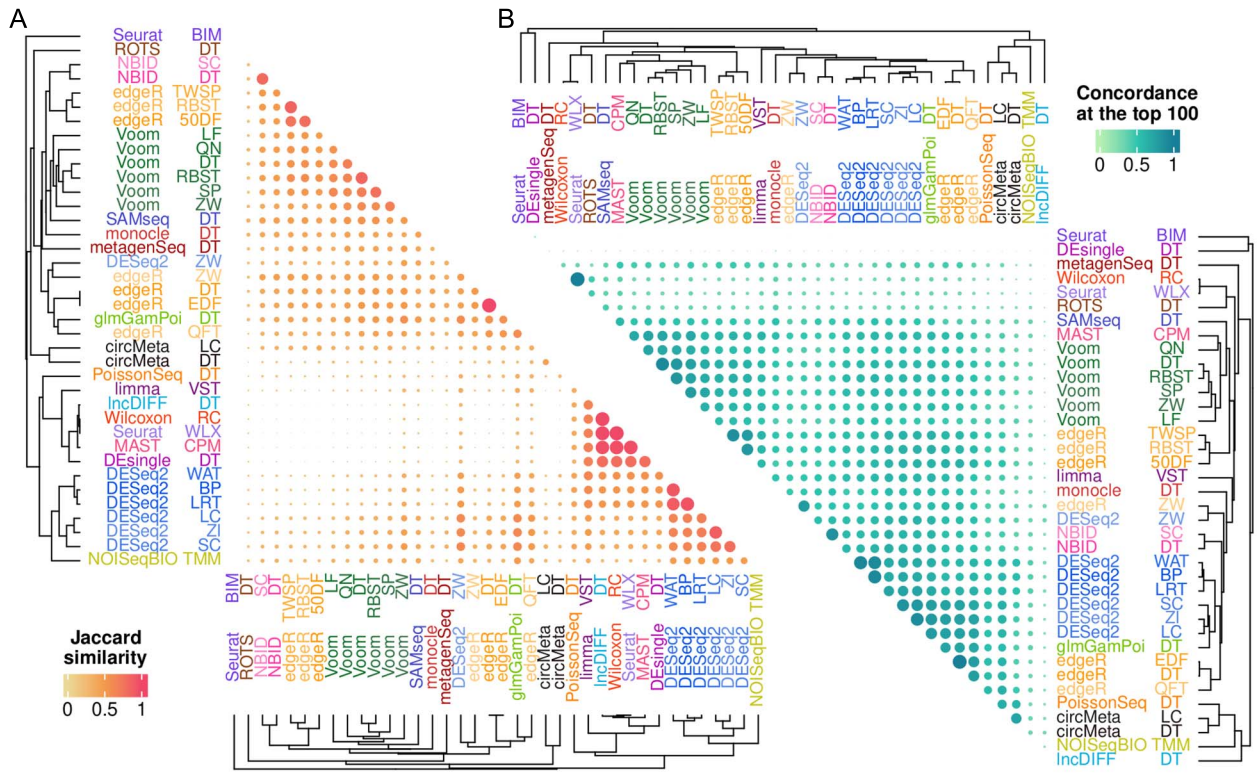

Figure 4. Similarity between the DEMs' predictions. (**A**) The dendrograms show the DEMs clustered according to (A) the Jaccard similarity for significantly DECs predicted at an adjusted *P*-value ≤0.05, and (**B**) the concordance of the 100 top circRNAs ranked according to *P*-values. The dot matrix shows the (A) Jaccard similarities and (B) concordance at the top between DEM pairs calculated as the average among all simulated data sets of 10 samples (N05). The dot size is proportional to the similarity scores.

Moreover, RNA-seq data are commonly purged of low-counted elements [31, 52, 53] to improve the performance of DEMs [15]. However, with circRNAs, a similar data filtering might not be desirable as we observed that removing lowly expressed circRNAs can discard most of the detected circRNAs, thus omitting much information.

Ad-hoc data filtering may be unnecessary with properly configured DEMs' parameters [20]. Therefore, we compared various parameter configurations of widely used bulk-RNA-seq tools. Plus, the large fraction of small and zero counts observed in circRNA data motivated us to evaluate metagenome and scRNA-seq DEMs, which, from a statistical point of view, can handle data with such characteristics [30]. For the first time and unlike previous studies [36], our analysis assessed the performance of scRNA-seq DEMs on bulk RNA-seq data.

We collected an unprecedented and extensive set of DEMs by selecting tools available as R packages, currently maintained and functioning, reasonably fast, and well-performing or never tested in previous benchmarks. We did not consider statistical models devised to assess the variation of the circular-to-linear expression ratio (CLR), such as CircTest [54], seekCRIT [55], DEBKS [56] and the circMeta test for CLR [22], because they address a problem distinct from differential circRNA abundance. Similarly, we did not include our recent method based on generalized linear mixed models [57] because it addresses the problem of combining circRNA expression quantification from multiple tools for differential expression analysis. We also considered two recent consensus models for scRNA-seq differential expression assessment [58, 59], but they failed to terminate the analysis on our data sets; finally, we assessed a new test from the convolution of multivariate hypergeometric distributions for differential expression [60], which performed poorly with circRNA data (data not shown).

We evaluated a few library size estimation strategies compatible with the selected DEMs and promising for circRNA data characteristics. In DESeq2, the *poscounts*, *shorth* and deconvolution functions to compute size factors showed better results than the default approach. In edgeR and NBID, the TMMwsp and deconvolution normalization procedures slightly improved the prediction of DECs. Finally, Limma-Voom worsened its FDP using quantile normalization compared with the TMM normalization. Nevertheless, a thorough comparison of normalization procedures as in previous works [61] was not our aim and warrants a dedicated study.

Our comparative study complies with best practices and tools for benchmarking bioinformatics methods [62–64]. CircRNA RNA-seq comparative experiments with many samples or with ground truth of differential expression are scarce or absent in public repositories. Therefore, we used a semi-parametric approach [40] to allow us to generate multiple data sets of different sample sizes. We focused on most typical scenarios of RNA-seq differential expression studies, thus limiting the maximum size of the data sets to 10 samples per group, as larger numbers of replicates are uncommon. Nevertheless, the design of our benchmark enabled us to observe a clear trend of the methods' performance upon increasing data set size. Besides, we obtained one pre-computed circRNA expression data set [47] with enough sample replicates to simulate unbiased data sets assuming no specific distribution underlying the expression data [48]. Although limited to one real data set, these non-parametric simulations mostly corroborated the results observed in the semi-parametric data.

Downloaded from https://academic.oup.com/bib/article/24/1/bbac612/6966517 by guest on 19 September 2024

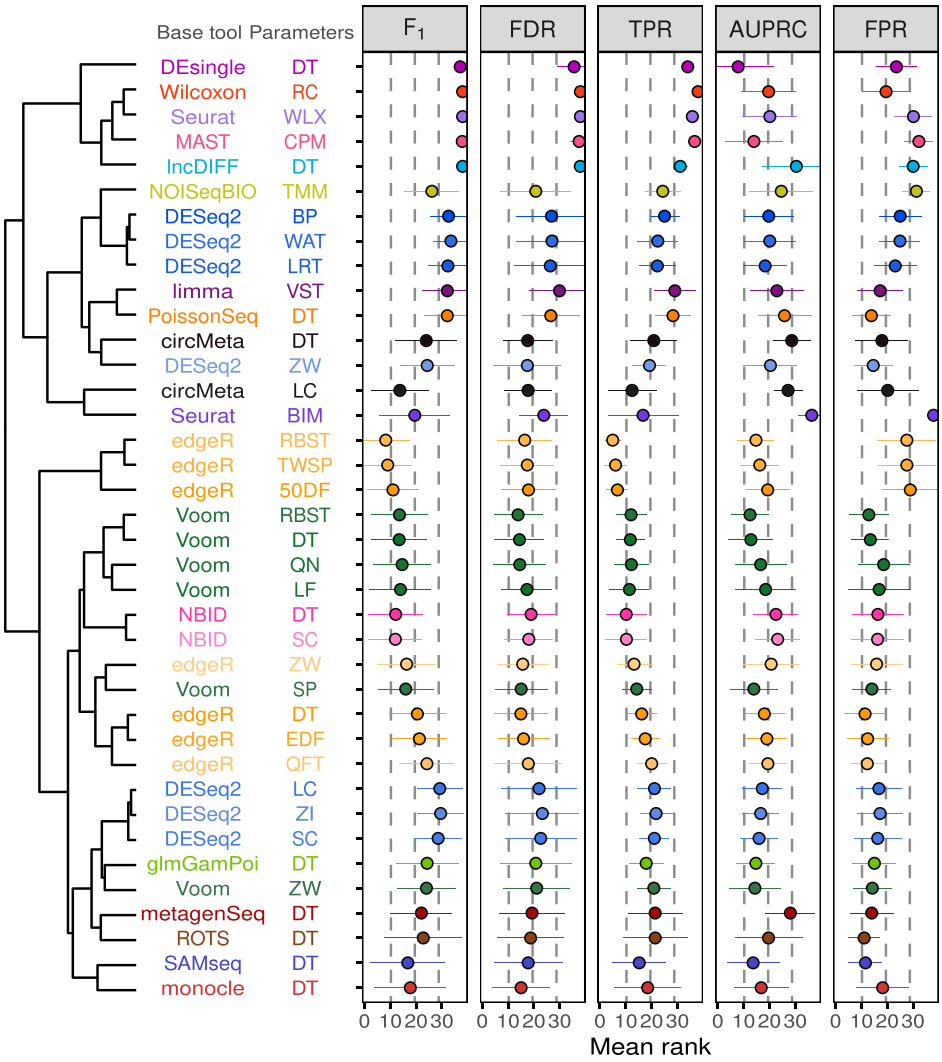

Figure 5. Overall ranking of the DEMs. Mean ranks (dots) and standard deviations (segments) of each method were computed on the simulated data sets with five replicates per condition group (N05) using four performance measures. Lower ranks correspond to better performances. Dashed lines indicate the first, median and third quartiles of ranks. The dendrogram on the left was computed with Canberra distance, considering the ranks of all the measures and using complete linkage. F_1: F_1 score; FDR: false discovery rate; TPR: true positive rate; AUC: area under the precision-recall curve; FPR: false positive rate.

A few aspects of the DEMs' performance were similar to the results of previous studies on low-count transcripts from bulk and scRNA-seq data [12, 13, 15–17, 20, 51, 65, 66].

We observed that most Limma-Voom pipelines controlled the type I error close to the nominal value, whereas DESeq2 showed a more conservative behaviour. Further, the parameter choice affected the type I error control in edgeR, showing a tendency towards higher FPR when imposing high degrees of freedom [16, 20].

Generally, the methods that controlled the FDR well showed low sensitivity; moreover, DEMs had higher power with larger data sets. DESeq2, edgeR, Limma and PoissonSeq confirmed poor sensitivity similarly to lowly counted transcripts in bulk-RNA-seq [13, 20], especially with a small number of replicates [13, 51]. The non-parametric DEMs, SAMseq and NOISeqBIO, required a higher replicate number to perform as well as other models [12] also in circRNA data. Moreover, we confirmed that SAMseq showed better FDR control than DESeq2 while retaining a high TPR [15,16]. Diversely, NOISeqBIO had unstable results depending on the set size: with three replicates, it showed high TPR [51], whereas, with five replicates, it obtained an FDR lower than nominal at the cost of a severe TPR loss [16]. Finally, the DEsingle's opposite AUPRC performance observed between our semi-parametric and non-parametric simulations was consistent with previous work [65], confirming that DEsingle's results are unstable and might depend on the data set.

The DEMs also presented performance diverging from previous benchmark works, supporting that circRNA expression data have different characteristics than linear transcripts from standard bulk RNA-seq, scRNA-seq and metagenome data. In particular, in our analysis, the *P*-value distributions of the most conservative DEMs displayed a smooth increase towards $P = 1$, suggesting that some systematic deviation from the theoretical null distributions of the test statistics occurred [43]. Instead, in low-counted lncRNA data, the DEMs presented conservative distributions with a spike near $p = 1$ [16], possibly due to the lncRNA's high variability. Moreover, DESeq2 did not obtain high AUC as in scRNA-seq [15,65], denoting DESeq2 yields overly conservative *P*-values in circRNA data.

Downloaded from https://academic.oup.com/bib/article/24/1/bbac612/6966517 by guest on 19 September 2024

Further, we did not observe the same bias in the type of genes preferentially called differentially expressed as for some DEMs in scRNA-seq data [15]. In particular, edgeR-QLF incorrectly called significant lowly expressed genes with many zeros in scRNA-seq, whereas, in our results, its FPs showed a higher expression and a lower fraction of zeros compared to TNs. Plus, in contrast with results from scRNA-seq data [15], the quasi-likelihood framework with edgeR effectively reduced the FPR in circRNA data. Conversely, the quasi-likelihood framework was unexpectedly detrimental with Limma-Voom, although the signal-to-noise statistics were similar to those obtained with edgeR-QLF.

In our results, DESeq2 with the Wald test did not show more liberal results than with the LRT, as observed by Raithel et al. [20] with low-counted genes. Such discordant results could be due to dependence on the data [17]. Unlike in Assefa et al. [16], PoissonSeq adequately controlled the type I error with many samples and showed higher TPR with larger sets. However, its high FDR suggests that a Poisson distribution does not fit circRNA expression data well.

Similarly to results from 16S and WMS data sets [17], MAST performed poorly at each sample size and obtained an overly conservative FPR, whereas, in scRNA-seq data, it performed reasonably well [15], also with lowly expressed isoforms [65]. MetagenomeSeq showed FPR below the nominal value in our analysis, whereas it obtained liberal FPR in 16S and WMS data [17], suggesting that metagenomeSeq has less consistent behaviour in data of different types. Monocle performed poorly in the smallest data sets of our benchmark but reached reasonable FDR and TPR with five or more samples per condition, which contrasted with previous analyses on scRNA-seq data [15,65]. Seurat-BIM reached a poor AUC, unlike in scRNA-seq data [15], suggesting it underperforms on small count data. Differently from our results, the Wilcoxon test performed reasonably on lowly expressed isoforms in scRNA-seq data [65]. Finally, the non-parametric methods clustered apart, differently from the results on scRNA-seq data [15].

We expected tools with specific options or designed for addressing small counts and sparse data, such as the scRNA-seq and metagenome methods, to outperform standard tools in circRNA data. Many scRNA-seq methods were clustered with bulk RNA-seq DEMs, primarily according to their underlying distribution model. Surprisingly, a few scRNA-seq pipelines, including Monocle, NBID, ROTS and glmGamPoi, performed reasonably well in circRNA data showing comparable or better predictions than bulk RNA-seq DEMs, especially in data sets of a large sample size.

Our work highlighted the challenging features of circRNA expression estimated with bulk RNA-seq. The generally poor performance of the methods could in part be explained by the conservative setting of our benchmark as the 'signal' data sets, containing a small proportion (10%) of DECs, can be regarded as a challenging scenario for the DEMs. We speculate that slightly better DEMs' performance would be obtained with higher fractions of DECs, as was observed in previous work for low-count lncRNAs [16]. Additionally, our comprehensive comparison of statistical tools for differential expression assessment applied to circRNA data marked a few caveats in circRNA expression analysis. We observed that no single method overperformed the others in every aspect and that the recall rate was generally low with RNA-seq data sets of typical size. Using default values in some methods can result in suboptimal performance. Conversely, custom parameter configurations can profoundly affect the predictions and contravene the expected performance. For instance, edgeR, one of the most used tools for RNA-seq data analysis, can provide misleading or poor predictions depending on its settings. EdgeR-RBST and edgeR-TWSP showed high FDR, whereas edgeR-QFT, although controlling FDR well, showed reduced TPR. Besides, Limma-Voom with default parameters controlled the type I error well and maintained a good trade-off between precision and recall but showed worse performance with the quasi-likelihood framework designed to better model zero counts. Instead, DESeq2, perhaps the most used tool in circRNA expression differential assessment, is overly conservative and underperforms several other methods, especially with data sets of typical size, no matter the parameters used. Notably, SAMseq, one of the oldest tools considered in our benchmark initially devised for microarray data and later adapted to RNA-seq, showed good results compared to its competitors. Further, scRNA-seq methods, such as glmGamPoi, monocle and ROTS, showed promising results in data sets with 20 replicates and could inspire novel solutions for circRNA data analysis.

Conclusions

This study shed light on the difference between circRNA and traditional gene expression RNA-seq data. RNA-seq studies willing to inspect circRNA expression require carefully balancing the trade-off between a higher sequencing depth and the number of replicates to obtain robust results. Our findings indicate that circRNA differential expression assessment from RNA-seq urges the development of new and robust computational models addressing the issues that emerged in our analysis. Our comprehensive benchmark highlights the importance of selecting an appropriate tool and configuring its parameters according to the data set characteristics and can guide biostatisticians and bioinformatics researchers in the analysis of circRNA differential expression.

Materials and methods

CircRNA data sets, expression quantification and expression filters

We analysed six independent data sets from circRNA studies (Table 1) for 235 samples in total. The JHS data set [67] considered Illumina sequencing data of 17 human tissues, with matched ribosomal RNA-depleted and ribosomal RNA-depleted followed by RNase R treatment libraries to enrich the circular transcript fraction. The other data sets (DM1 [68], IDC [69], IPF [70], MS [71] and PC [47]) considered ribosomal RNA-depleted Illumina sequencing libraries of tumour and relative healthy tissue samples, with at least five biological replicates per group and more than 40 million sequenced paired-end reads.

Raw reads of the JHS, DM1, IDC, IPF and MS data sets were available in public data sets such as SRA and NGDC, whereas the circRNA expression data of the PC data set were provided by the authors of the original study [47]. The PC data set consisted of 29,234 circRNAs from 31 normal tissues and 126 prostate cancer samples. The BJR counts were computed with CIRI2, discarding circRNAs detected with less than two BJRs; further details are available in the original article [47].

CircRNAs of the JHS, DM1, IDC, IPF and MS data sets were detected and quantified with CirComPara2 v0.1.2.1 [26] using default parameters. CirComPara2 runs seven circRNA identification pipelines and combines their results to obtain reliable detections and expression quantification. In addition, from the CirComPara2 output files, we obtained the circRNAs quantified by CIRI2 [72], CIRCexplorer2 [73] v2.3.8 using either TopHat-Fusion

Downloaded from https://academic.oup.com/bib/article/24/1/bbac612/6966517 by guest on 19 September 2024

v2.1.0 or Segemehl v0.3.4, DCC [54] v0.4.8 and Findcirc [74] v1.2. Moreover, we applied CIRIquant [75] v1.1.2, giving as input the circRNAs detected with CirComPara2 to obtain circRNA expression abundance also with this tool. Note that these methods encompass five different read aligners (Bowtie2, BWA-MEM, Segemehl, STAR and TopHat-Fusion), plus one re-alignment method based on BWA-MEM, thus limiting possible biases derived from the mapping algorithms.

We explored the circRNA expression of each comparison data set by performing the principal component analysis (PCA) on the CirComPara2 expression estimates and removing circRNAs detected in less than three samples. PCA plots showed circRNA expression patterns associated with the main groups, indicating significant variation of circRNA expression between conditions (Supplementary Figure S25).

Five strategies to discard circRNAs were applied independently to the circRNA expression matrices. Each circRNA must be detected in (i) any sample (unfiltered data), (ii) at least three samples, (iii) at least half the samples, (iv) all the biological replicates of at least one group (to keep only the circRNAs consistently expressed within a condition) and (v) at least as many samples as the size of the smallest group (such a filter can be helpful when groups have largely different numbers of replicates).

Details of the goodness-of-fit analysis are reported in Supplementary Material.

Semi-parametric simulations

The SPsimSeq R package v1.4.0 [40] was used to simulate data sets from real data. SPsimSeq uses the Gaussian-copulas to retain the between-genes correlation structure and allows generating of arbitrarily large data sets. The original data sets underwent preliminary processing that considered the removal of circRNAs expressed in less than three samples, followed by selecting samples with a similar library size in both the sample groups. No samples were discarded for IDC and IPF data sets; the DM1 data set resulted in five control and six tumour samples, and the MS data set resulted in 12 control and 18 tumour samples. For each original data set, 30 simulations were run considering two sample groups of an equal number of samples, with set sizes of 6, 10 and 20 samples and unvaried library sizes between simulations. The number of circRNAs simulated corresponded to the circRNAs detected in the original data sets after the quality filters: 1490 in DM1, 7540 in IDC, 2485 in IPF and 7217 in MS. Two types of data sets were generated: 'null' data sets, where no differentially expressed circRNAs were simulated, and 'signal' data sets, where 10% circRNAs were significantly differentially expressed between the sample groups and with an absolute log-fold-change ≥ 0.5. Each simulated data set underwent a preliminary filter to remove the simulated circRNAs with non-zero counts in less than three samples.

The quality metrics of the simulated data sets were computed with the countsimQC package's functions [41] custom optimized for parallel execution.

Non-parametric simulations

We performed a preliminary quality assessment of the PC data set to determine sample batches via inspection of the first two principal components (Supplementary Figure S26). Then, we selected samples of only one batch to obtain homogeneous samples, for a total of 8239 circRNAs from 20 normal tissues and 76 prostate cancer samples.

We performed non-parametric simulations with the SimSeq tool [48], which does not impose any distribution assumption on the simulated data. We generated 90 instances for 'null' and 'signal' data sets with 6, 10 and 20 samples of two equal replicate number condition groups.

Differential expression tools used in this study

In our analysis, we considered the following differential expression tools and package versions: CircMeta v1.0.2 [22], DESeq2 v1.22.2 [23], DEsingle v1.14.0 [76], edgeR v3.36.0 [24, 31, 34], glmGamPoi v1.6.0 [77], Limma v3.50.0 [78], Limma-Voom v3.50.0 [35], lncDIFF v1.0.0 [79], MAST v1.20.0 [80], metagenomeSeq v1.36.0 [81], Monocle v2.22.0 [82], NBID v0.1.2 [33], NOISeqBIO v2.38.0 [83], PoissonSeq v1.1.2 [84], ROTS v1.22.0 [85], SAMSeq v3.0 [86], Seurat v4.1.0 [87, 88], the Wilcoxon test, ZINB-Wave v1.16.0, genefilter v1.76.0, scran v1.22.1 and sctransform v0.3.3. Further details on the parameter configurations used in the differential expression pipelines are reported in Supplementary Methods.

For all algorithms, the *P*-values from genes with a non-zero sum of read counts across samples were adjusted using the Benjamini–Hochberg procedure [89].

Type I error control

For this analysis, we used the 'null' simulated data set without differentially expressed circRNAs. The *P*-values returned by each method were used to compare the number of false discoveries upon thresholds of 0.1, 0.05 and 0.01. For NOISeqBIO, we used its scores in place of *P*-values only for completeness, but they were not considered for comparison with other methods.

Concordance at the top

We used the concordance at the top (CAT) to evaluate concordance for each DEM. Starting from two lists of ranked features by *P*-values, the CAT statistic was computed in the following way. For a given integer i, concordance is defined as the cardinality of the intersection of the top i elements of each list, divided by i, i.e. $\#\{L_{1:i} \cap M_{1:i}\}/i$, where L and M represent the two lists. This concordance was computed for values of i from 1 to R.

Depending on the study, only a minority of features may be expected to be differentially expressed between two experimental conditions. Hence, the expected number of differentially expressed features is a good choice as the maximum rank R. The CAT displays high variability for low ranks as few features are involved, whereas concordance tends to 1 as R approaches the total number of features, becoming uninformative. We set $R = 100$, considering this number biologically relevant and high enough to permit an accurate concordance evaluation. We used CAT for Between-Method Concordance (BMC), in which a method is compared to other methods in the same simulated data set to evaluate consistency. To summarise this information for all pairwise method comparisons, we computed the area under the curve, giving a better score to the method pairs consistently concordant for all values of i from 1 to R.

Additional software used in this study

The featureCounts function [90] from the Rsubread v2.8.1 package [91] was used to compute read alignment counts. AUPRC were computed with the PRROC v1.3.1 R package [92] using the Davis & Goadrich algorithm [46]. Plots were generated using ggplot2 v3.3.5. The computational time of differential expression tools was measured with the tictoc v1.0.1 R package. All simulation analyses were run using the SummarizedBenchmark v2.12.0 framework [63].

Downloaded from https://academic.oup.com/bib/article/24/1/bbac612/6966517 by guest on 19 September 2024

Key Points

- CircRNA expression RNA-seq data is characterized by a high proportion of small counts and zeros.
- Traditional methods for differential expression assessment underperform when applied to circRNA expression data, especially in small-size data sets.
- Specific parameter configurations can improve differential expression methods' performance on circRNA expression analysis.
- Differential expression tools devised for single-cell RNA-seq data analysis perform reasonably with circRNA expression data.
- Differential expression methods perform best in data sets with large number of replicates.

Supplementary Data

Supplementary data are available online at https://academic.oup.com/bib.

Data availability

The RNA-seq data sets used in this study can be accessed from the NGDC repository [93] with accession number PRJCA000751 (for JHS data); the GEO repository [94] with accession numbers GSE86356 (for the DM1 data), GSE52463 (for the IPF data) and GSE159225 (for the MS data); and the SRA repository [29] with accession number SRP156355 (for the IDC data); the PC data are available on request from the authors of [47].

The code used in this work is available at https://github.com/egaffo/DEM4circ.

Authors' contributions

A.B.: conceptualization, data curation, formal analysis, methodology, investigation, software, visualization, writing—original draft, writing—review and editing. S.B.: funding acquisition, project administration, resources, supervision, writing—review and editing. E.G.: conceptualization, data curation, formal analysis, methodology, investigation, software, visualization, project administration, resources, supervision, validation, writing—original draft, writing—review and editing.

Funding

Ministero dell'Istruzione, dell'Università e della Ricerca (PRIN 2017 #2017PPS2X4_003 to S.B.); Associazione Italiana per la Ricerca sul Cancro, Milan, Italy (Investigator Grant 2017 #20052 to S.B.); EU funding within the MUR PNRR "National Center for Gene Therapy and Drugs based on RNA Technology" (Project no. CN00000041 CN3 Spoke #6 "RNA chemistry") and "National Center for HPC, Big Data and Quantum Computing" (Project no. CN00000013 CN1 Spoke #8 "In Silico Medicine & Omics Data").

References

1. Liu C-X, Chen L-L. Circular RNAs: characterization, cellular roles, and applications. *Cell* 2022;**185**(12):2016–34. https://doi.org/10.1016/j.cell.2022.04.021.
2. Buratin A, Paganin M, Gaffo E, *et al.* Large-scale circular RNA deregulation in T-ALL: unlocking unique ectopic expression of molecular subtypes. *Blood Adv* 2020;**4**:5902–14.
3. Dal Molin A, Hofmans M, Gaffo E, *et al.* CircRNAs dysregulated in juvenile myelomonocytic leukemia: CircMCTP1 stands out. *Front Cell Dev Biol* 2020;**8**:613540.
4. Kristensen LS, Jakobsen T, Hager H, *et al.* The emerging roles of circRNAs in cancer and oncology. *Nat Rev Clin Oncol* 2022;**19**:188–206.
5. Chen L, Wang C, Sun H, *et al.* The bioinformatics toolbox for circRNA discovery and analysis. *Brief Bioinform* 2021;**22**:1706–28.
6. An O, Tan K-T, Li Y, *et al.* CSI NGS portal: An online platform for automated NGS data analysis and sharing. *Int J Mol Sci* 2020;**21**(11):3828. https://doi.org/10.3390/ijms21113828.
7. Yu H, Jiao B, Lu L, *et al.* NetMiner-an ensemble pipeline for building genome-wide and high-quality gene co-expression network using massive-scale RNA-seq samples. *PLoS One* 2018;**13**:e0192613.
8. Gokool A, Loy CT, Halliday GM, *et al.* Circular RNAs: the brain transcriptome comes full circle. *Trends Neurosci* 2020;**43**:752–66.
9. Hua JT, Chen S, He HH. Landscape of noncoding RNA in prostate cancer. *Trends Genet* 2019;**35**:840–51.
10. Hansen TB. Improved circRNA identification by combining prediction algorithms. *Front Cell Dev Biol* 2018;**6**:20.
11. Anders S, McCarthy DJ, Chen Y, *et al.* Count-based differential expression analysis of RNA sequencing data using R and Bioconductor. *Nat Protoc* 2013;**8**:1765–86.
12. Soneson C, Delorenzi M. A comparison of methods for differential expression analysis of RNA-seq data. *BMC Bioinform* 2013;**14**.
13. Rapaport F, Khanin R, Liang Y, *et al.* Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data. *Genome Biol* 2013;**14**:R95.
14. Seyednasrollah F, Laiho A, Elo LL. Comparison of software packages for detecting differential expression in RNA-seq studies. *Brief Bioinform* 2015;**16**:59–70.
15. Soneson C, Robinson MD. Bias, robustness and scalability in single-cell differential expression analysis. *Nat Methods* 2018;**15**:255–61.
16. Assefa AT, De Paepe K, Everaert C, *et al.* Differential gene expression analysis tools exhibit substandard performance for long non-coding RNA-sequencing data. *Genome Biol* 2018;**19**:96.
17. Calgaro M, Romualdi C, Waldron L, *et al.* Assessment of statistical methods from single cell, bulk RNA-seq, and metagenomics applied to microbiome data. *Genome Biol* 2020;**21**:191.
18. Xu C, Zhang J. Mammalian circular RNAs result largely from splicing errors. *Cell Rep* 2021;**36**:109439.
19. Szabo L, Salzman J. Detecting circular RNAs: bioinformatic and experimental challenges. *Nat Rev Genet* 2016;**17**:679–92.
20. Raithel S, Johnson L, Galliart M, *et al.* Inferential considerations for low-count RNA-seq transcripts: a case study on the dominant prairie grass Andropogon gerardii. *BMC Genomics* 2016;**17**:140.
21. Warton DI. Why you cannot transform your way out of trouble for small counts. *Biometrics* 2018;**74**:362–8.
22. Chen L, Wang F, Bruggeman EC, *et al.* circMeta: a unified computational framework for genomic feature annotation and differential expression analysis of circular RNAs. *Bioinformatics* 2020;**36**:539–45.
23. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* 2014;**15**:550.

Downloaded from https://academic.oup.com/bib/article/24/1/bbac612/6966517 by guest on 19 September 2024

24. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 2010;**26**:139–40.
25. Hansen TB, Venø MT, Damgaard CK, *et al.* Comparison of circular RNA prediction tools. *Nucleic Acids Res* 2016;**44**:e58.
26. Gaffo E, Buratin A, Dal Molin A, *et al.* Sensitive, reliable and robust circRNA detection from RNA-seq with CirComPara2. *Brief Bioinform* 2022;**23**(1):bbab418. https://doi.org/10.1093/bib/bbab418.
27. Nielsen AF, Bindereif A, Bozzoni I, *et al.* Best practice standards for circular RNA research. *Nat Methods* 2022;1–13.
28. Soneson C, Love MI, Robinson MD. Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. *F1000Res* 2015;**4**:1521.
29. Leinonen R, Sugawara H, Shumway M, *et al.* The sequence read archive. *Nucleic Acids Res* 2011;**39**:D19–21.
30. Jiang R, Sun T, Song D, *et al.* Statistics or biology: the zero-inflation controversy about scRNA-seq data. *Genome Biol* 2022;**23**:31.
31. Chen Y, Lun ATL, Smyth GK. From reads to genes to pathways: differential expression analysis of RNA-Seq experiments using Rsubread and the edgeR quasi-likelihood pipeline. *F1000Res* 2016;**5**:1438.
32. Conesa A, Madrigal P, Tarazona S, *et al.* A survey of best practices for RNA-seq data analysis. *Genome Biol* 2016;**17**.
33. Chen W, Li Y, Easton J, *et al.* UMI-count modeling and differential expression analysis for single-cell RNA sequencing. *Genome Biol* 2018;**19**:70.
34. Zhou X, Lindsay H, Robinson MD. Robustly detecting differential expression in RNA sequencing data using observation weights. *Nucleic Acids Res* 2014;**42**:e91.
35. Law CW, Chen Y, Shi W, *et al.* Voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol* 2014;**15**:R29.
36. Van den Berge K, Perraudeau F, Soneson C, *et al.* Observation weights unlock bulk RNA-seq tools for zero inflation and single-cell applications. *Genome Biol* 2018;**19**:24.
37. Risso D, Perraudeau F, Gribkova S, *et al.* A general and flexible method for signal extraction from single-cell RNA-seq data. *Nat Commun* 2018;**9**:284.
38. Lun ATL, Bach K, Marioni JC. Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. *Genome Biol* 2016;**17**:75.
39. Hafemeister C, Satija R. Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. *Genome Biol* 2019;**20**:296.
40. Assefa AT, Vandesompele J, Thas O. SPsimSeq: semi-parametric simulation of bulk and single-cell RNA-sequencing data. *Bioinformatics* 2020;**36**:3276–8.
41. Soneson C, Robinson MD. Towards unified quality verification of synthetic count data with countsimQC. *Bioinformatics* 2018;**34**:691–2.
42. Lun ATL, Smyth GK. No counts, no variance: allowing for loss of degrees of freedom when assessing biological variability from RNA-seq data. *Stat Appl Genet Mol Biol* 2017;**16**:83–93.
43. Breheny P, Stromberg A, Lambert J. P-value histograms: inference and diagnostics. *High-Throughput* 2018;**7**:23.
44. Zhou X, Robinson MD. Do count-based differential expression methods perform poorly when genes are expressed in only one condition? *Genome Biol* 2015;**16**.
45. Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS One* 2015;**10**:e0118432.
46. Davis J, Goadrich M. The relationship between Precision-Recall and ROC curves. *Proceedings of the 23rd International Conference on Machine Learning - ICML '06*. New York, NY, USA: Association for Computing Machinery, 2006, 233–240. https://doi.org/10.1145/1143844.1143874.
47. Hansen EB, Fredsøe J, Okholm TLH, *et al.* The transcriptional landscape and biomarker potential of circular RNAs in prostate cancer. *Genome Med* 2022;**14**:8.
48. Benidt S, Nettleton D. SimSeq: a nonparametric approach to simulation of RNA-sequence datasets. *Bioinformatics* 2015;**31**:2131–40.
49. Bartlett MS. The use of transformations. *Biometrics* 1947;**3**:39.
50. Aufiero S, Reckman YJ, Tijsen AJ, *et al.* circRNAprofiler: an R-based computational framework for the downstream analysis of circular RNAs. *BMC Bioinform* 2020;**21**:164.
51. Schurch NJ, Schofield P, Gierliński M, *et al.* How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use? *RNA* 2016;**22**:839–51.
52. Rau A, Gallopin M, Celeux G, *et al.* Data-based filtering for replicated high-throughput transcriptome sequencing experiments. *Bioinformatics* 2013;**29**:2146–52.
53. Bourgon R, Gentleman R, Huber W. Independent filtering increases detection power for high-throughput experiments. *Proc Natl Acad Sci* 2010;**107**:9546–51.
54. Cheng J, Metge F, Dieterich C. Specific identification and quantification of circular RNAs from sequencing data. *Bioinformatics* 2016;**32**:1094–6.
55. Chaabane M, Andreeva K, Hwang JY, *et al.* seekCRIT: detecting and characterizing differentially expressed circular RNAs using high-throughput sequencing data. *PLoS Comput Biol* 2020;**16**:e1008338.
56. Liu Z, Ding H, She J, *et al.* DEBKS: a tool to detect differentially expressed circular RNA. *Genom Proteom Bioinform* 2021;**20**(3):549–56. https://doi.org/10.1016/j.gpb.2021.01.003.
57. Buratin A, Romualdi C, Bortoluzzi S, *et al.* Detecting differentially expressed circular RNAs from multiple quantification methods using a generalized linear mixed model. *Comput Struct Biotechnol J* 2022;**20**:2495–502.
58. Li H-S, Ou-Yang L, Zhu Y, *et al.* scDEA: differential expression analysis in single-cell RNA-sequencing data via ensemble learning. *Brief Bioinform* 2022;**23**.
59. Zou J, Deng F, Wang M, *et al.* scCODE: an R package for data-specific differentially expressed gene detection on single-cell RNA-sequencing data. *Brief Bioinform* 2022;**23**.
60. Tumminello M, Bertolazzi G, Sottile G, *et al.* A multivariate statistical test for differential expression analysis. *Sci Rep* 2022;**12**:1–10.
61. Bullard JH, Purdom E, Hansen KD, *et al.* Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinformatics* 2010;**11**:94.
62. Weber LM, Saelens W, Cannoodt R, *et al.* Essential guidelines for computational method benchmarking. *Genome Biol* 2019;**20**:125.
63. Kimes PK, Reyes A. Reproducible and replicable comparisons using SummarizedBenchmark. *Bioinformatics* 2019;**35**:137–9.
64. Soneson C, Robinson MD. *iCOBRA: Open, Reproducible, Standardized and Live Method Benchmarking*.
65. Mou T, Deng W, Gu F, *et al.* Reproducibility of methods to detect differentially expressed genes from single-cell RNA sequencing. *Front Genet* 2019;**10**:1331.
66. Stupnikov A, McInerney CE, Savage KI, *et al.* Robustness of differential gene expression analysis of RNA-seq. *Comput Struct Biotechnol J* 2021;**19**:3470–81.

Downloaded from https://academic.oup.com/bib/article/24/1/bbac612/6966517 by guest on 19 September 2024

67. Ji P, Wu W, Chen S, *et al.* Expanded expression landscape and prioritization of circular RNAs in mammals. *Cell Rep* 2019;**26**:3444–3460.e5.
68. Wang ET, Treacy D, Eichinger K, *et al.* Transcriptome alterations in myotonic dystrophy skeletal muscle and heart. *Hum Mol Genet* 2019;**28**:1312–21.
69. Rao AKDM, Arvinden VR, Ramasamy D, *et al.* Identification of novel dysregulated circular RNAs in early-stage breast cancer. *J Cell Mol Med* 2021;**25**:3912–21.
70. Nance T, Smith KS, Anaya V, *et al.* Transcriptome analysis reveals differential splicing events in IPF lung tissue. *PLoS One* 2014;**9**:e97550.
71. Iparraguirre L, Alberro A, Sepúlveda L, *et al.* RNA-Seq profiling of leukocytes reveals a sex-dependent global circular RNA upregulation in multiple sclerosis and 6 candidate biomarkers. *Hum Mol Genet* 2020;**29**:3361–72.
72. Gao Y, Zhang J, Zhao F. Circular RNA identification based on multiple seed matching. *Brief Bioinform* 2018;**19**:803–10.
73. Zhang X-O, Dong R, Zhang Y, *et al.* Diverse alternative backsplicing and alternative splicing landscape of circular RNAs. *Genome Res* 2016;**26**:1277–87.
74. Memczak S, Jens M, Elefsinioti A, *et al.* Circular RNAs are a large class of animal RNAs with regulatory potency. *Nature* 2013;**495**:333–8.
75. Zhang J, Chen S, Yang J, *et al.* Accurate quantification of circular RNAs identifies extensive circular isoform switching events. *Nat Commun* 2020;**11**:90.
76. Miao Z, Deng K, Wang X, *et al.* DEsingle for detecting three types of differential expression in single-cell RNA-seq data. *Bioinformatics* 2018;**34**:3223–4.
77. Ahlmann-Eltze C, Huber W. glmGamPoi: fitting Gamma-Poisson generalized linear models on single cell count data. *Bioinformatics* 2021;**36**:5701–2.
78. Ritchie ME, Phipson B, Wu D, *et al.* Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res* 2015;**43**:e47.
79. Li Q, Yu X, Chaudhary R, *et al.* lncDIFF: a novel quasi-likelihood method for differential expression analysis of non-coding RNA. *BMC Genomics* 2019;**20**:539.
80. Finak G, McDavid A, Yajima M, *et al.* MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biol* 2015;**16**.
81. Paulson JN, Stine OC, Bravo HC, *et al.* Differential abundance analysis for microbial marker-gene surveys. *Nat Methods* 2013;**10**:1200–2.
82. Trapnell C, Cacchiarelli D, Grimsby J, *et al.* The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat Biotechnol* 2014;**32**:381–6.
83. Tarazona S, Furió-Tarí P, Turrà D, *et al.* Data quality aware analysis of differential expression in RNA-seq with NOISeq R/Bioc package. *Nucleic Acids Res* 2015;**43**:e140.
84. Li J, Witten DM, Johnstone IM, *et al.* Normalization, testing, and false discovery rate estimation for RNA-sequencing data. *Biostatistics* 2012;**13**:523–38.
85. Suomi T, Seyednasrollah F, Jaakkola MK, *et al.* ROTS: An R package for reproducibility-optimized statistical testing. *PLoS Comput Biol* 2017;**13**:e1005562.
86. Li J, Tibshirani R. Finding consistent patterns: a nonparametric approach for identifying differential expression in RNA-Seq data. *Stat Methods Med Res* 2013;**22**:519–36.
87. Hao Y, Hao S, Andersen-Nissen E, *et al.* Integrated analysis of multimodal single-cell data. *Cell* 2021;**184**:3573–3587.e29.
88. Satija R, Farrell JA, Gennert D, *et al.* Spatial reconstruction of single-cell gene expression data. *Nat Biotechnol* 2015;**33**:495–502.
89. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc B Methodol* 1995;**57**:289–300.
90. Liao Y, Smyth GK, Shi W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* 2014;**30**:923–30.
91. Liao Y, Smyth GK, Shi W. The R package Rsubread is easier, faster, cheaper and better for alignment and quantification of RNA sequencing reads. *Nucleic Acids Res* 2019;**47**:e47–7.
92. Grau J, Grosse I, Keilwagen J. PRROC: computing and visualizing precision-recall and receiver operating characteristic curves in R. *Bioinformatics* 2015;**31**:2595–7.
93. Members CNCB-NGDC, Partners XY, Bao Y, *et al.* Database resources of the National Genomics Data Center, China National Center for Bioinformation in 2022. *Nucleic Acids Res* 2021;**50**:D27–38.
94. Edgar R. Gene expression omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res* 2002;**30**:207–10.

Downloaded from https://academic.oup.com/bib/article/24/1/bbac612/6966517 by guest on 19 September 2024