Computational and Structural Biotechnology Journal 20 (2022) 2495–2502

COMPUTATIONAL ANDSTRUCTURAL BIOTECHNOLOGY JOURNAL

journal homepage: www.elsevier.com/locate/csbj

Detecting differentially expressed circular RNAs from multiple quantification methods using a generalized linear mixed model

Alessia Buratin [a,b], Chiara Romualdi [b], Stefania Bortoluzzi [a,c,1], Enrico Gaffo [a,*,1]

[a] Department of Molecular Medicine, University of Padova, Padova, Italy
[b] Department of Biology, University of Padova, Padova, Italy
[c] Interdepartmental Research Center for Innovative Biotechnologies (CRIBI), University of Padova, Padova, Italy

ARTICLE INFO

Article history:
Received 14 April 2022
Received in revised form 11 May 2022
Accepted 12 May 2022
Available online 20 May 2022

Keywords:
Circular RNAs
circRNAs
Differential expression
Generalized linear mixed models
RNA-seq

ABSTRACT

Finding differentially expressed circular RNAs (circRNAs) is instrumental to understanding the molecular basis of phenotypic variation between conditions linked to circRNA-involving mechanisms. To date, several methods have been developed to identify circRNAs, and combining multiple tools is becoming an established approach to improve the detection rate and robustness of results in circRNA studies. However, when using a consensus strategy, it is unclear how circRNA expression estimates should be considered and integrated into downstream analysis, such as differential expression assessment. This work presents a novel solution to test circRNA differential expression using quantifications of multiple algorithms simultaneously. Our approach analyzes multiple tools' circRNA abundance count data within a single framework by leveraging generalized linear mixed models (GLMM), which account for the sample correlation structure within and between the quantification tools. We compared the GLMM approach with three widely used differential expression models, showing its higher sensitivity in detecting and efficiently ranking significant differentially expressed circRNAs. Our strategy is the first to consider combined estimates of multiple circRNA quantification methods, and we propose it as a powerful model to improve circRNA differential expression analysis.

1. Introduction

Circular RNAs (circRNAs) are a large class of RNA molecules in which a downstream splice donor site is covalently closed to an upstream acceptor site by an event called backsplicing. Several circRNAs have been proven to govern cellular processes [7], and circRNA dysregulation can drive disease and cancer mechanisms [16]. Thus, the characterization and quantification of circRNAs from high-throughput RNA-seq data have become the subject of many studies. Moreover, identifying the differentially expressed circRNAs is a prerequisite to detecting circRNAs potentially involved in disease mechanisms and prioritizing those requiring further functional investigation.

Most bioinformatics tools developed to detect circRNAs from RNA-seq data quantify circRNA expression by counting the back spliced reads at each back splicing junction [3]. Recent strategies can improve the circRNA recall rate by concurrently applying multiple detection methods [6,4]. However, different circRNA detection methods give different expression estimates, and it is not always straightforward to combine their outputs to obtain a single quantification measure for each circRNA in each sample, a data structure that is required for most of the downstream analysis. Of note, differential circRNA expression testing performed with traditional count-based models [10,14,9] imposes the use of single expression estimates and cannot model quantifications from multiple algorithms. Given that single circRNA detection tools suffer from low detection sensitivity, such an approach might artificially limit the discovery of differentially-expressed circRNAs (DECs) because the quantification method overlooked or was unable to correctly estimate the abundance of some DECs. Ideally, all reliable

Abbreviations: AUC, Area under the ROC curve; circRNAs, circular RNAs; DECs, Differentially Expressed circRNAs; DEMs, Differential Expression Models; FDR, False Discovery Rate; GLMM, Generalized Linear Mixed Model; RNAseq, RNA sequencing; TPR, True Positive Rate.

* Corresponding author.

E-mail addresses: alessia.buratin@studenti.unipd.it (A. Buratin), chiara.romualdi@unipd.it (C. Romualdi), stefania.bortoluzzi@unipd.it (S. Bortoluzzi), enrico.gaffo@unipd.it (E. Gaffo).

[1] Equally last authors.

https://doi.org/10.1016/j.csbj.2022.05.026
2001-0370/© 2022 The Author(s). Published by Elsevier B.V. on behalf of Research Network of Computational and Structural Biotechnology.

circRNAs detected should be tested for differential expression, but no method currently allows such an analysis.

In this work, taking advantage of the improved detection rate and repeated measures given by the expression estimates of different models, we propose a new method based on generalized linear mixed models (GLMM) to assess circRNA differential expression. Several metrics were used to evaluate the GLMM approach's performance, including the type I error control, power, replicability across datasets, and internal consistency among methods.

2. Results and discussion

2.1. A generalized linear mixed model for modeling multiple circRNA expression estimates

We propose a novel approach for circRNA differential expression analyses that exploits multiple methods' estimates to improve statistical power.

In previous work, we showed that different tools might output substantially dissimilar sets of circRNAs and quantification estimates [4]. CircRNA expression quantification obtained with different tools from the same data can be considered pseudo-repeated measures. Under this assumption, we expect that measures quantified by the same detection tool (intra-method correlation) would be more correlated with each other than those quantified by different tools (inter-method correlation), thus entailing a hierarchical structure. We evaluated this assumption by comparing the distribution of the pairwise correlations of the circRNA expression estimated by (i) the same method (intra-method correlation), and (ii) pairs of different methods (inter-method correlation) using three datasets (Fig. 1) and four circRNA quantification methods. Our results confirmed that the intra-method was higher than the inter-method correlation supporting the use of the generalized mixed model.

We used GLMMs to account for the correlation among samples measured with different detection tools, including the chosen detection tools as random effects. Since expression values are count data, we modeled circRNA expression with a negative binomial distribution (NB) embedded into the GLMM. GLMMs can easily be adapted to assess differential expression between conditions through a likelihood ratio test. Further details on GLMMs can be found in the "Materials and Methods" section.

2.2. Comparison with commonly used differential expression tools

We compared the GLMM approach with widely-used tools for RNA-seq differential expression analysis, namely DESeq2 [10], edgeR [11], and Limma-Voom [9], on simulated and real circRNA data sets. The methods were evaluated in terms of power, type I error control, replicability across datasets, and internal consistency (see "Materials and Methods").

The GLMM approach considers expression estimates of multiple tools at once, whereas other differential expression methods (DEMs) analyze only single-method estimates. Consequently, the set of circRNAs analyzed with the GLMM approach might be larger than the set considered by other DEMs. Such an unbalance in the number of circRNAs favors the GLMM approach in terms of statistical power. To overcome this bias, each DEM was applied separately to each expression matrix (one for each detection tool), obtaining as many lists of P-values as the number of detection tools. For each circRNA, we selected the lowest non-null value among the (detection tool) P-value lists, obtaining, for each DEM, a unique P-value list that included all circRNAs considered by the GLMM approach. Note that by choosing the lowest P-values, we expect higher type I error and false discovery rates but higher sensitivity for the DEMs. Finally, each unique P-value list underwent correction for multiple tests.

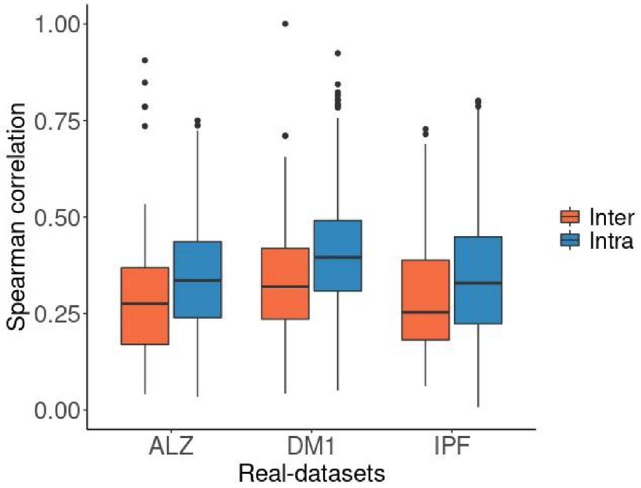

Fig. 1. Sample correlation between and within circRNAs quantification methods. The circRNA expression was estimated independently with four quantification methods, and the pairwise Spearman's correlations were calculated between different (Inter, red boxes) or the same (Intra, blue boxes) quantification method circRNA expression estimates in real RNA-seq datasets of three independent human circRNA studies (x-axis). ALZ: brain tissue, Alzheimer's disease; DM1: skeletal muscle tissue, Myotonic Dystrophy Type 1; IPF: lung tissue, Idiopathic Pulmonary Fibrosis. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

2.2.1. Parametric simulations

We simulated 30 circRNA expression data using a NB distribution with parameters estimated from the ALZ dataset (Supplementary Table 1). The sample size (10 samples), the proportion of DECs (10%), and the condition effect size (1.5 fold change) were maintained constant. The GLMM approach and each DEM were then applied to the 30 simulated datasets.

We assessed the algorithms' true positive rate (TPR), their control of the false-positive calls, and the area under the ROC curve (AUC). The performances of the methods are presented in Table 1. The GLMM approach achieved the highest power but showed an overly conservative control of the false discovery rate (FDR), which was lower than the nominal level of 0.05. Limma-voom obtained the second highest TPR but with a significantly larger FDR. DESeq2 and edgeR better used their FDR budget but showed less power. Finally, the GLMM approach presented the highest AUC, indicating its better ranking of DECs according to P-values.

2.2.2. Real-data benchmark

We used experimental reproducibility on independent samples to obtain ground truth approximation, following the scheme previously presented in another benchmark study [10]. Briefly, we repeatedly split each dataset into one evaluation and one verification set, assigning a larger sample size to the verification set. Then, we compared the circRNAs called as significant from the two sets, holding the verification set as the ground truth (Supplementary Fig. 2). We used the ALZ and IPF datasets in which we expect truly differentially expressed circRNAs exist, whereas the DM1 dataset did not have enough sample replicates in its groups to perform the non-parametric simulations.

2.2.2.1. Type I error control. We evaluated each algorithm's type I error rate control, i.e., the probability of the statistical test calling a feature significantly differentially expressed when it is not. We simulated expression profiles under the null hypothesis of no dif-

Table 1
Summary of simulation results. Average Area under ROC curve (AUC), True Positive Rate (TPR), Sensitivity and False Discovery Rate (FDR) across 30 simulated datasets from ALZ data are reported for each Differential Expression Method (DEM).

DEM	TPR	FDR	AUC
DESeq2	0.416	0.043	0.944
edgeR	0.591	0.066	0.785
limma-voom	0.614	0.101	0.796
GLMM	0.683	0.016	0.998

ferential expression, randomly assigning samples to a group and repeating this process 30 times. In this setting, the observed proportion of falsely significant tests at the observed α should match the nominal value (e.g., α = 0.05). In the context of differential expression analysis, liberal tests will lead to many false discoveries, while conservative tests will control the type I error at the cost of reduced power, potentially hindering true discoveries.

We observed that DESeq2, edgeR, and limma-voom had an overly conservative control of the type I error, whereas the GLMM approach obtained an observed α closer to the nominal value. The low type I error of the other DEMs was in contrast with what we expected from the construction of the P-value list of each DEM. However, this could be explained by the small number of replicates and low expression levels considered in the real-dataset simulations, as observed in previous work [13]. This result suggests that the GLMM approach benefits from the additional information provided by the multiple expression estimates.

2.2.2.2. Sensitivity and precision. To compute the sensitivity and precision of the algorithms, we considered the recombination of the original samples of the ALZ and IPF datasets (Supplementary Fig. 2). Moreover, we considered as the reference truth the predictions on the verification set of one algorithm at a time and compared them with each method's calls on the evaluation set to evaluate their performance. In doing so, we could compare each method against (i) itself by comparing its predictions on the verification and evaluation sets, and (ii) all other methods by comparing its predictions on the verification set with other methods' predictions on the evaluation set. We used this approach rather than a consensus-based method, as we did not want to favor or disfavor any particular algorithm or group of algorithms. Sensitivity and precision were calculated by setting an adjusted P-value ≤ 0.1 significance threshold (Fig. 2).

Fig. 3 displays the estimates of sensitivity and precision for each algorithm pair using the two datasets. The ranking of algorithms was generally consistent regardless of which algorithm was chosen to compose the verification set. We observed neatly different results in the two datasets. In the ALZ data set, which comprises more samples per group and a high sequencing depth, the median sensitivity estimates were typically between 0 and 0.2 for all methods (Fig. 3a). In both datasets, all algorithms had a relatively low median sensitivity that can be explained by the small sample size of the evaluation set and the fact that increasing the sample size in the verification set increases the power.

The precision estimates are displayed in Fig. 3b, where the highest median precision was often reached by GLMM. In general, we observed a high precision in both datasets across the different algorithms used as the gold standard. To inspect the ability of GLMM to effectively reduce the number of false negatives (FNs), Supplementary Fig. 3 provides an intersection plot of DECs for one random replication of ALZ datasets. We observed, in this example, the ability of GLMM to recognize the larger overlap of calls within each other DEMs. In addition, the absolute number of calls for the evaluation and verification sets can be seen in Supplementary Fig. 4, which mostly matched the order seen in the sensitivity plot of Fig. 3, highlighting the capacity of the proposed method to detect in both evaluation and verification sets the higher number of calls.

Notably, sensitivity and precision analysis results on real-data benchmark confirmed observations of the parametric simulations (Table 1). Moreover, although the classical models were very con-

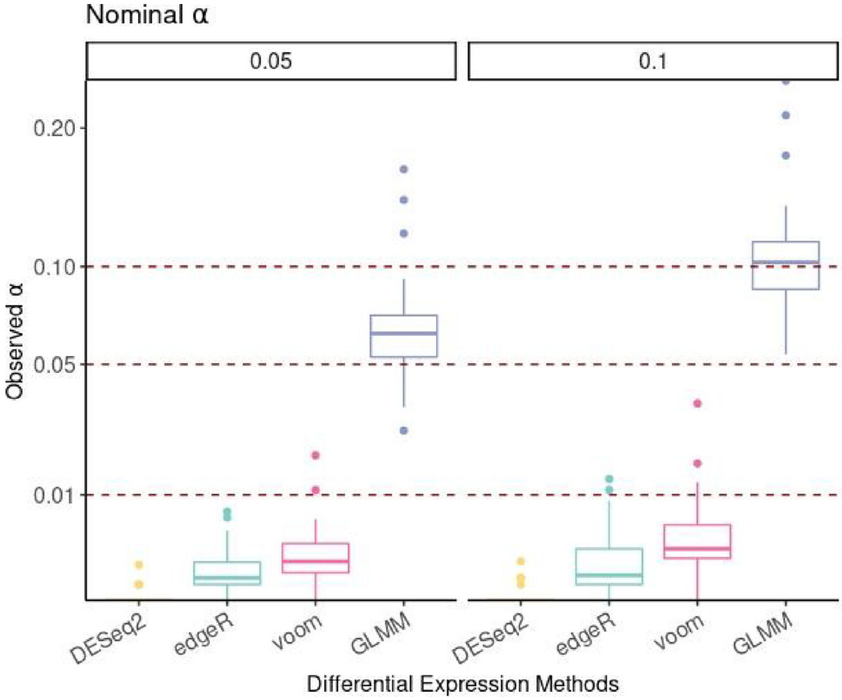

Fig. 2. Type I error control. Boxplots of the proportion of tests with P-values lower than nominal αs (0.05 and 0.1) in 30 mock datasets obtained from ALZ data samples. The dashed red line indicates 0.01, 0.05, and 0.1 type I error thresholds. The y-axis was squared-root scaled to improve visibility. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

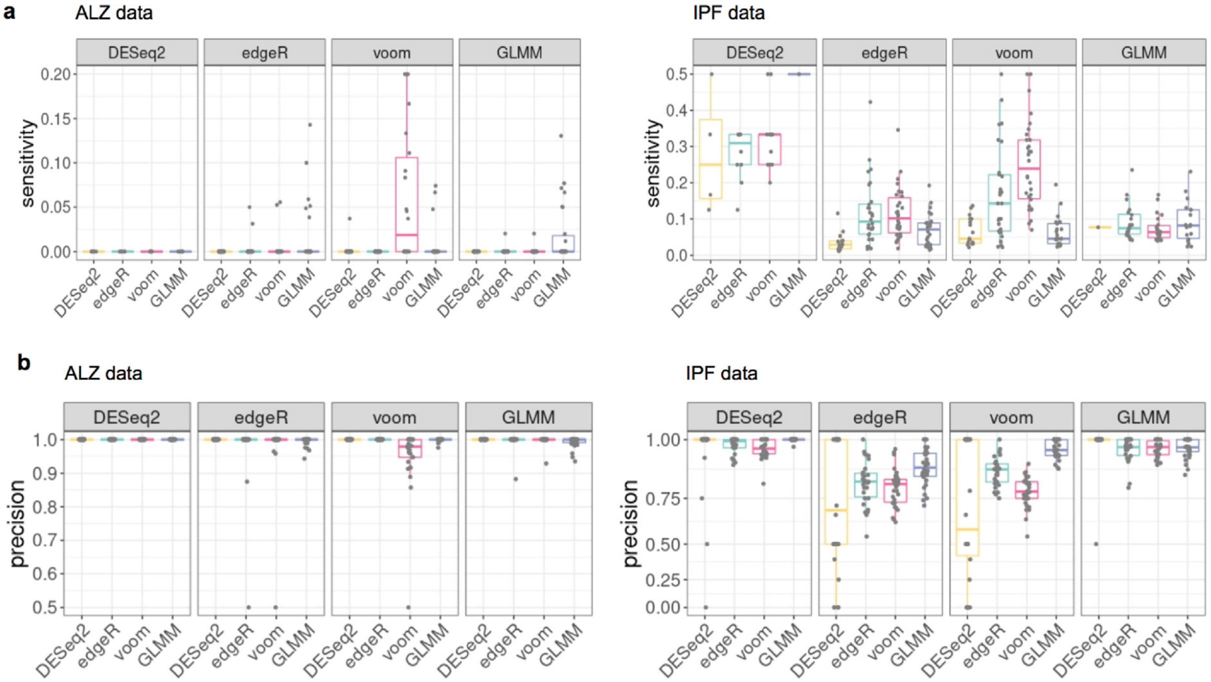

Fig. 3. Sensitivity and Precision of real-data benchmark. Sensitivity (a) and precision (b) of each algorithm (in the horizontal axis) when holding one algorithm's predictions on the verification set as the reference truths (in the facet strip labels) and comparing them with the algorithms' predictions on the evaluation sets.

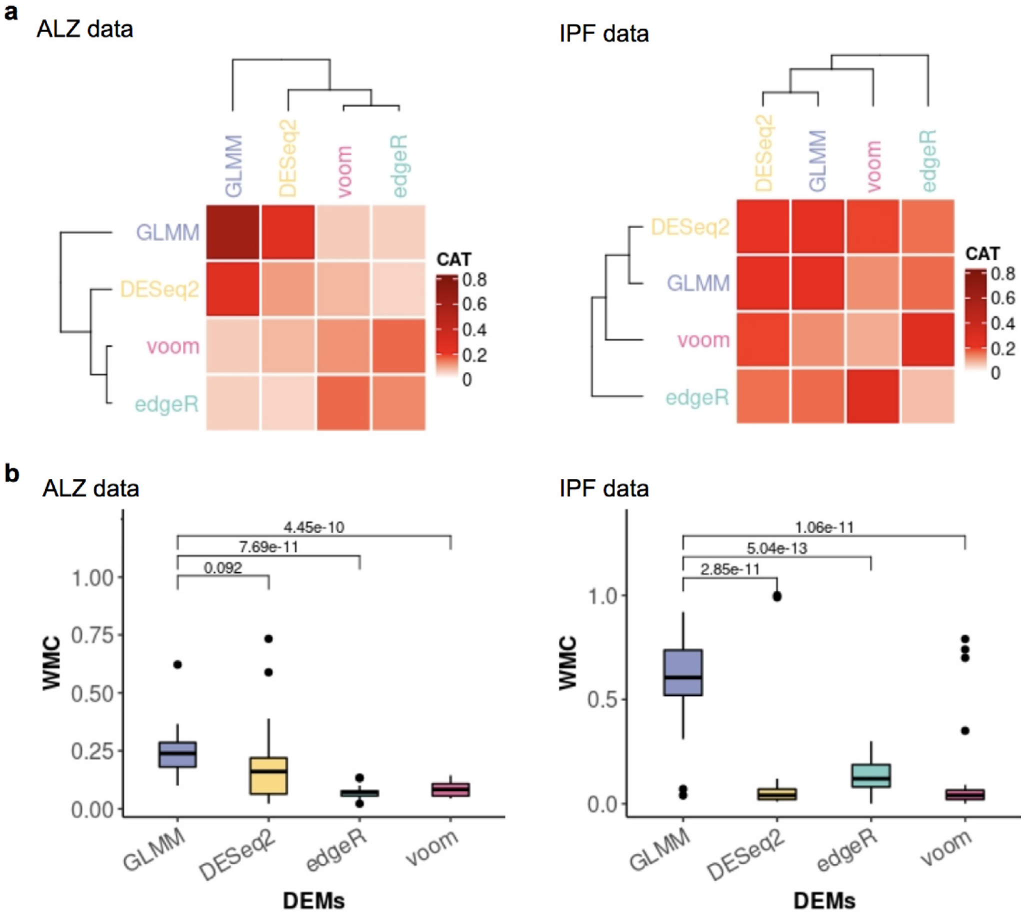

Fig. 4. Consistency and replicability of models. (a) Average between-method (non-diagonal cells) and within-method (main diagonal cells) concordance for the 100 top-ranked calls DEMs in replicated ALZ and IPF datasets. (b) Boxplot of the within-method concordance (WMC) on ALZ and IPF data. The plot illustrates that the concordances are different by comparing each other with the GLMM model in terms of WMC. P-value denotes the result from a pairwise *t*-test. CAT: concordance at the 100 top differentially expressed circRNAs.

servative in type I error control (Fig. 2), we demonstrate the ability of the GLMM to reduce the rate of false positives supported by higher precision, detecting fewer false negative DECs in comparison to the classical DEMs commonly used. These results indicate that by using the GLMM, we can gain sensitivity and, at the same time, derive much more information from the data, even when the sample size and the sequencing depth are limited.

2.2.2.3. Concordance and replicability. To measure the ability of each method to produce concordant results between methods and replicable results in independent data, we used the ALZ and IPF datasets. Samples were split into two subsets, and each DEM was separately applied to each subset. The process was repeated 30 times.

The concordance-at-the-top (CAT) metric was used for assessing the within-method concordance (WMC) and between-method concordance (BMC). We used the BMC to (i) group methods based on their degree of agreement and (ii) identify those methods sharing the largest amount of discoveries with the majority of the other methods. Consistency is required for method validity: methods sharing most DECs with the majority of other methods are more likely to produce valid results.

Concordance analysis performed on the ALZ dataset showed that the methods clustered within two groups: the first comprising the GLMM and the second containing the classical GLM models (Fig. 4a). A different picture emerged from the analysis of the IPF dataset, where GLMM clustered with DESeq2, whereas voom and edgeR had similar results as previously observed in the ALZ dataset. In general, we note that GLMM has the potential to reach a higher BMC in both datasets compared to other DEMs, in particular when compared to the DESeq2 model.

Real-data benchmarks could be useful to validate differential expression findings. To account for this, we used the CAT metric to assess the WMC, i.e., the amount of concordance of the results of each method on the two random evaluation and verification sets.

The highest WMC was obtained with the GLMM method (Fig. 4b), indicating its significantly higher consistency of predictions in both datasets.

In summary, the benchmarking tests showed that GLMM effectively controlled type-I errors, maintaining a median false positive rate just below the chosen critical value in a mock comparison of groups of samples randomly chosen from a larger pool. For both simulation and analysis of real data, GLMM often achieved the highest sensitivity of those algorithms that controlled the FDR.

3. Conclusions

We investigated different theoretical and practical issues related to the analysis of circRNA data. The main objective of our study was to improve differential circRNA expression assessment by exploiting data from multiple circRNA quantification tools and explore a solution to the problem of the multiple circRNA expression estimates arising from the consensus detection strategies that are being increasingly used in circRNA studies.

In three independent RNA-seq data of circRNA studies, we observed that the within-detection-tool correlation was larger than between-detection tools (Fig. 1), encouraging the use of mixed models as a way to account for pseudo-repeated observations. Overall, mixed effect models lead to the most accurate results when analyzing data with a correlation structure [18].

The lack of real circRNA datasets with a ground truth makes the assessment of DECs challenging. However, we implemented an assessment strategy to obtain a reasonably good evaluation of methods' performances in terms of false discoveries, replicability, and sensitivity.

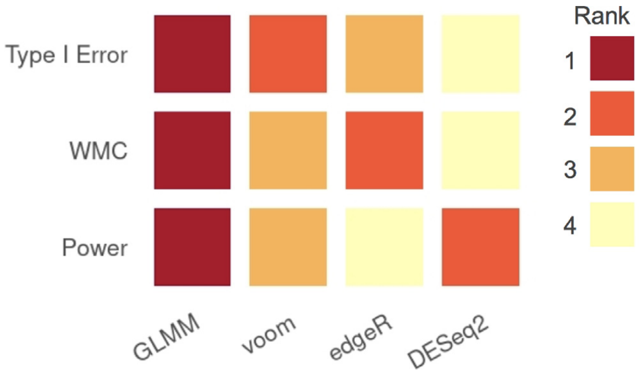

Fig. 5. Ranking of the methods based on three evaluation criteria. The type I error ranking was based on the analysis of the 30 mock comparisons from ALZ datasets; the within method concordance (WMC) was based on the average WMC values across the 30 random subset comparisons for each of the two datasets used (ALZ and IPF); the power analysis ranking was based on the tumor vs. normal ALZ and IPF dataset evaluations. The ranks range from 1 (best) to 4 (worst) with lower rank values corresponding to better performances.

We have shown that the GLMM approach increases the DEC detection and robustness by leveraging outputs from different circRNA quantification tools. In this study, we used four top-performing circRNA detection and quantification methods [6,4]. However, our model is not limited to the tools we applied in our analysis; instead, it can accept any combination of tools for estimating circRNA expression according to the user's needs and choices, as long as expression data is provided with unprocessed read counts. Notwithstanding this, as a general rule in circRNA expression quantification, we recommend selecting combinations of tools that provide reliable predictions [4].

We compared the GLMM approach to widely used DEMs in two datasets with different features and showed its consistent higher performance, as summarized in Fig. 5, in which the methods were ranked according to the different performance metrics. Our power simulations (Table 1) and type I error (Fig. 2) suggest that important true effects could be retrieved from the combined circRNA expression matrix. The decreased power observed is due to an overestimation of the mean-square error relative to mixed-effects models, particularly when the intra-methods variance is larger than the inter-methods variance, as appears to be typical with circRNA data (Fig. 1).

Although our focus here is on testing for differential circRNA expression between two conditions, GLMMs are flexible enough for considering other complex designs and allow diverse correlation structures with random effects. However, similarly to any other differential expression method applied to bulk RNA-seq data, our GLMM approach does not prevent the potential problems associated with performing differential gene expression analysis using complex tissues consisting of many different cell types [8], and the GLMM predictions should be cautiously interpreted.

In conclusion, the GLMM approach is advantageous in the identification of differentially expressed circRNAs even if the number of samples is limited, as demonstrated in real data and simulation benchmarks.

4. Materials and methods

4.1. Real data sets and circRNA expression quantification

All datasets are publicly available. A complete list of the datasets' main characteristics is reported in Supplementary Table 1.

Backsplice junction read counts were estimated using four among the most used and top-performing methods [6,4]: CIRI2 v2.0.6 [5], DCC v0.4.8 [2], findcirc v1.2 [12], and CircExplorer2 v2.3.8 [17]. These tools implement different strategies for backsplice detection and circRNA expression quantification, overall involving three read aligner algorithms, such as BWA, Bowtie2, and STAR (see Supplementary Methods for parameters' details).

4.2. Implementation of the generalized linear mixed model approach

Starting from circRNA count matrices obtained from different detection tools, we constructed a combined count matrix composed of the count estimates of each detection tool for each sample. Specifically, having count matrices with circRNAs in rows and samples in columns, the combined matrix will result in as many rows as the set of the circRNAs detected by at least two methods and the number of columns as large as the number of samples multiplied by the number of the quantification methods (Supplementary Fig. 1). The circRNAs identified by only one method are in general less reliable [6,4] and were therefore excluded.

Let y_{cij} ($c = 1,...,C$; $i = 1,...,n$; $j = 1,...,m$), the raw count measurements of circRNA *c* for the *i*th sample, obtained by the *j*th detection method, and x_i a known p-dimensional vector with the covariate information corresponding to the *i*th row of the $n \times p$ model matrix X_i, known as the design matrix, which contains the values of multiple sample characteristics. In particular, in X_i both time-varying (e.g. measurement time, environmental conditions) and time-independent covariates (e.g. treatment group, baseline age, gender, etc.) are allowed. We used a q dimensional vector of random effects b_i to model the serial correlation with the corresponding design matrix Z_i. For a specific circRNA *c*, the generalized mixed-effects model assumes $Y_{cij}|b_{ci} \sim NB(\mu_{cij},\phi_c)$, where ϕ_c represents the circRNA-wise dispersion parameter that measures overdispersion. Based on this parameterization, $E(Y_{ijc}|b_{ci}) = \mu_{ijc}$ is modeled as a function of the explanatory variables x_{ij}, random effects b_{ci}, and an offset term α_{ij}. In particular, we considered, for a single circRNA *c*, the following:

$$g\left(\mu_{ijc}\right) = \alpha_{ij} + x^{T}{}_{ij}\beta_c + Z_i b_{ic}, \tag{1}$$

where α_{ij} is an offset term with the logarithm of the effective library size derived from edgeR. In this work, we used the trimmed mean method (TMM) of Robinson et al. [15] to calculate the scaling factors to correct for sequencing depth and potentially composition bias. The group effects β are assumed to follow a normal distribution $N(\beta_0 1_p, \sigma^2{}_\beta I_p)$, where 1_p is the $p \times 1$ dimensional vector whose elements are all 1, I_p is the $p \times p$ dimensional identity matrix, β_0 and $\sigma^2{}_\beta$ are mean and variance of the normal distribution and *p* is the number of covariates. The quantification method variable $\mathbf{Z}_i$ introduces hierarchical dependence of circRNA counts and is included in the model as a random effect design matrix. We assume that the quantification method random effects follow a multivariate normal distribution $\mathbf{b}_{ci} \sim N_q(0,\Psi_c)$, where Ψ_c is a positive-definite variance–covariance matrix that determines the form and complexity of the random effects. GLMMs under NB distributions are considered to model y_{ij}.

For inference on fixed-effects β, the null hypothesis $H_0{:}\beta = 0$ is tested against the alternative $H_1{:}\beta \neq 0$ with a likelihood-ratio test (LRT). The random effects b can be tested by z-statistic for difference from 0.

4.3. Intra- and inter-correlation analysis

For each dataset, we made pairwise comparisons to compute intra- and inter-method correlations of circRNAs expression estimates. To control for the correlation structure between circRNAs, we subsequently trimmed from the dataset the circRNAs with Spearman's correlation coefficient > 0.25. This step was repeated until either no more uncorrelated circRNAs remained or a total of 500 uncorrelated circRNAs were obtained. The intra-methods Spearman's correlation was computed for all possible pairs of samples measured with the same quantification method. The inter-method Spearman's correlation was computed for all possible pairs of samples, randomly drawing ten times two samples measured with different detection tools.

4.4. Differential expression tools and analysis approach

We applied three widely used tools for assessing differential expression in RNA-seq studies, namely DESeq2 (v1.34.0), edgeR (v3.36.0), and limma-voom (v3.38.3) [9]. DESeq2 and edgeR fit negative binomial distributions for count data with generalized linear models, whereas limma-voom uses normalized counts in logarithmic scale to estimate a mean–variance relationship and compute appropriate observational-level weights with a linear model. The linear model's residual degrees of freedom were adjusted before the empirical Bayes variance shrinkage and were propagated to the moderated statistical tests. DESeq2 and limma-voom were used with default parameters. EdgeR considered TMM normalization, tagwise robust dispersion estimation, and quasi-likelihood F test.

We analyzed the circRNA expression estimated by four detection tools, i.e., CIRI2, CIRCexplorer, DCC, and findcirc, using three differential expression models (DEMs): DESeq2, edgeR, and limma-voom. Each DEM was independently applied to the expression matrices estimated by the circRNA detection tools, obtaining four lists of P-values per DEM. For each DEM, we merged the P-value lists by selecting the lowest non-null P-value obtained for each circRNA *c* detected by at least two DEMs:

$$P - value_{DEM}(c) = min\{P - value_{DEM,CIRI2}(c), P - value_{DEM,CIRCexp}(c), P - value_{DEM,DCC}(c), P - value_{DEM,findcirc}(c)\}$$

Afterward, multiple test corrections were applied to the merged list using the Benjamini–Hochberg procedure [1].

4.5. GLMM

The glmmTBM package (v1.1.1) fits a generalized linear mixed model using Template Model Builder (TMB), using automatic differentiation to estimate model gradients and the Laplace approximation for handling random effects. A glmmTMB model has four main components: a conditional model formula, a distribution for the conditional model, and a dispersion model formula. Simple GLMMs can be fitted using the conditional model while dispersion formulas at their default values. The mean of the conditional model is specified using a two-sided formula with the response variable on the left and predictors on the right, potentially including random effects and offsets. In our analysis, we used count ~ group + (1 | detection_tool) to evaluate if circRNA counts vary by condition (group) and vary randomly by detection tool used for their quantification. We examined the NB, using *family="nbinom2"*. A likelihood ratio test was then used to assess the significance of circRNA differential expression between sample groups, and P-values were corrected for multiple testing by using the Benjamini–Hochberg (BH) procedure [1].

4.6. Power analysis

We used sensitivity and "1 - specificity" to evaluate the power of DE methods. We simulated datasets of 5,000 circRNAs with negative binomial distributed counts. To simulate data with realistic moments, the mean and dispersions were drawn from the joint distribution of means and gene-wise dispersion estimates, fitting only an intercept term, from the real dataset obtained using four detection tools (Supplementary Materials). Datasets with a total sample size of 10 were simulated, and the samples were split into two equal-sized groups; 90% of the simulated circRNAs had no true differential expression, while for 10% of the circRNAs, a true fold change of 1.5 was used to generate counts across the two groups, with the direction of fold change chosen randomly. The simulated differentially expressed circRNAs were chosen uniformly at random among all the circRNAs, throughout the range of mean counts.

4.7. Real data benchmark

A random split of the original datasets into an evaluation and a verification subset was replicated 30 times (Supplementary Fig. 2). In particular, the ALZ dataset, which contains 9 human disease samples compared to 8 samples of their normal counterparts, allowed Evaluation sets of 3 vs. 3 and Verification sets of 6 vs. 5 samples.

For a given algorithm's verification set calls, we tested the evaluation set call of each other algorithm included in the comparison and of the same algorithm (each algorithm with itself). CircRNA as 'true' differentially expressed was defined by an adjusted P-value < 0.1 in the larger verification set.

Note that the calls from the verification set are only an approximation of the true differential state, and the approximation error has one systematic and one stochastic component. The stochastic error becomes small once the sample size of the verification set is large enough. For the systematic errors, our benchmark assumes that these affect all algorithms equally and do not markedly change the ranking of the algorithms.

4.7.1. Type I error control

For this analysis, we used the collection of ALZ samples. We randomly split the samples into two groups of 6 and 8 samples, respectively. We repeated the random split 30 times and applied the DEMs to each split dataset. Every method returned a p-value for each feature. Those values were used to compare the number of false discoveries with the common thresholds of 0.05 and 0.1.

4.7.2. Sensitivity and precision

The sensitivity was calculated as the fraction of circRNAs with true differences between group means, with true differential expression defined by an adjusted P-value ≤ 0.1 in the verification set. The precision was calculated as the fraction of circRNAs true positives in the set of those passing the adjusted P-value threshold. This can also be reported as $1 - \text{FDR}$.

4.7.3. Replicability and consistency

We used the Concordance At the Top (CAT) measure to evaluate the concordance of different differential expression methods. 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., #{L1:i ∩ M1: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. In fact, CAT displays high variability for low ranks as few features are involved, while 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 Within-Method Concordance (WMC) and Between-Method Concordance (BMC). In the first one, a method is compared to itself in random splits of the datasets to assess the replicability, whereas in the BMC, a method is compared to other methods in the same dataset to evaluate consistency.

To evaluate the WMC, for each algorithm, the list of features ordered by p-values obtained from the evaluation set was compared to those obtained in the verification set; whereas to evaluate the BMC, the list of features ordered by p-value obtained from the evaluation set was compared to the analogous list obtained from evaluation set by all the other DEMs. WMC and BMC were averaged across the 30 replicates to obtain the final values.

4.8. Source code availability

The source code used in this work is available at https://github.com/AFBuratin/DECMiMo.

Funding

This work has been supported by Italian Ministry of Education, Universities and Research (PRIN 2017 #2017PPS2X4_003 to SB); AIRC, Milano, Italy (IG 2017 #20052 to SB, IG 2018 #21837 to CR).

CRediT authorship contribution statement

Alessia Buratin: Conceptualization, Data curation, Formal analysis, Methodology, Investigation, Software, Visualization, Writing – original draft, Writing – review & editing. **Chiara Romualdi:** Methodology, Supervision, Validation, Writing – review & editing. **Stefania Bortoluzzi:** Funding acquisition, Project administration, Resources, Supervision, Writing – review & editing. **Enrico Gaffo:** Conceptualization, Data curation, Methodology, Investigation, Project administration, Supervision, Validation, Writing – original draft, Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We would like to thank Prof. Monica Chiogna from the University of Bologna for her help in the revision of the manuscript.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.csbj.2022.05.026.

References

[1] Benjamini Y, Hochberg Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. J Roy Stat Soc: Ser B (Methodol) 1995. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x.

[2] Cheng J, Metge F, Dieterich C. Specific identification and quantification of circular RNAs from sequencing data. Bioinformatics 2016. https://doi.org/10.1093/bioinformatics/btv656.

[3] Chen L, Wang C, Sun H, Wang J, Liang Y, Wang Y, et al. The Bioinformatics toolbox for circRNA discovery and analysis. Briefings Bioinf 2021;22 (2):1706–28.

[4] Gaffo E, Buratin A, Dal Molin A, Bortoluzzi S. Sensitive, reliable and robust circRNA detection from RNA-Seq with CirComPara2. Brief Bioinf 2022;23(1). https://doi.org/10.1093/bib/bbab418.

[5] Gao Y, Zhang J, Zhao F. Circular RNA identification based on multiple seed matching. Brief Bioinf 2018. https://doi.org/10.1093/bib/bbx014.

[6] Hansen TB. Improved circRNA identification by combining prediction algorithms. Front Cell Dev Biol 2018;6(March):20.

[7] Kristensen LS, Andersen MS, Stagsted LVW, Ebbesen KK, Hansen TB, Kjems J. The biogenesis, biology and characterization of circular RNAs. Nat Rev Genet 2019;20(11):675–91.

[8] Kristensen LS, Ebbesen KK, Sokol M, Jakobsen T, Korsgaard U, Eriksen AC, et al. Spatial expression analyses of the putative oncogene ciRS-7 in cancer reshape the microRNA sponge theory. Nat Commun 2020;11(1):4551.

[9] Law CW, Chen Y, Shi W, Smyth GK. Voom: precision weights unlock linear model analysis tools for RNA-Seq read counts. Genome Biol 2014;15(2):R29.

[10] Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-Seq data with DESeq2. Genome Biol 2014;15(12):550.

[11] McCarthy DJ, Chen Y, Smyth GK. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. Nucleic Acids Res 2012;40(10):4288–97.

[12] Memczak S, Jens M, Elefsinioti A, Torti F, Krueger J, Rybak A, et al. Circular RNAs are a large class of animal RNAs with regulatory potency. Nature 2013. https://doi.org/10.1038/nature11928.

[13] Rapaport F, Khanin R, Liang Y, Pirun M, Krek A, Zumbo P, et al. Comprehensive evaluation of differential gene expression analysis methods for RNA-Seq data. Genome Biol 2013. https://doi.org/10.1186/gb-2013-14-9-r95.

[14] Robinson MD, McCarthy DJ, Smyth GK. edgeR: A bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 2010;26(1):139–40.

[15] Robinson MD, Oshlack A. A scaling normalization method for differential expression analysis of RNA-Seq data. Genome Biol 2010;11(3):R25.

[16] Tang X, Ren H, Guo M, Qian J, Yang Y, Chunyan G. Review on circular RNAs and new insights into their roles in cancer. Comput Struct Biotechnol J 2021;19 (January):910–28.

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

[18] Zimmerman KD, Espeland MA, Langefeld CD. A practical solution to pseudoreplication bias in single-cell studies. Nat Commun 2021;12(1):738.