© 2023. Published by The Company of Biologists Ltd | Journal of Cell Science (2023) 136, jcs260312. doi:10.1242/jcs.260312

TOOLS AND RESOURCES

Prog-Plot – a visual method to determine functional relationships for false discovery rate regression methods

Nicolás Bello* and Liliana López-Kleine

ABSTRACT

Multiple test corrections are a fundamental step in the analysis of differentially expressed genes, as the number of tests performed would otherwise inflate the false discovery rate (FDR). Recent methods for *P*-value correction involve a regression model in order to include covariates that are informative of the power of the test. Here, we present Progressive proportions plot (Prog-Plot), a visual tool to identify the functional relationship between the covariate and the proportion of *P*-values consistent with the null hypothesis. The relationship between the proportion of *P*-values and the covariate to be included is needed, but there are no available tools to verify it. The approach presented here aims at having an objective way to specify regression models instead of relying on prior knowledge.

KEY WORDS: RNA-Seq, Differential expression, False discovery rate, Genomics

INTRODUCTION

High-throughput transcriptome sequencing (RNA-seq) is a widely used technique to understand biological systems and, in general, phenotypic variation. Specifically, the detection of differentially expressed (DE) genes is a topic of constant discussion and improvement (see Costa-Silva et al., 2017).

An important part of the differential expression analysis is the multiple testing correction, given that the number of simultaneous tests usually varies in the range of thousands to tens of thousands. Although the correction proposed by Benjamin and Hochberg (1995) is still arguably the most used, there have been some new proposals to improve this step it in the last decade (see, for example, Korthauer et al., 2019).

Some of the most recent methods mainly include improvements related to power. One of them consists of including a regression step with an informative covariate (see Boca and Leek, 2018; Lei and Fithian, 2018 and Scott et al., 2015). All these methods rely on the relationship of the covariate to the *P*-value outcome of the test used for detecting differential expression. However, to the best of our knowledge, there is no systematic method to identify the functional relationship that these tools require. Until now, the researcher would need to have prior information on the informativeness of the covariate or achieve a satisfactory model through trial and error.

An important part of the correction proposed by Boca and Leek is the tuning parameter lambda (λ). This parameter is used in the estimation of the proportion of null *P*-values as it is shown in the Materials and Methods. The choice of a single value of λ constitutes a trade-off between bias and variance. A small value of λ will give biased estimates with low variance, and a value closer to 1 will give estimates with a smaller bias but that is highly variable. Like in other previous work, Boca and Leek tackle this issue by calculating estimates for a range of values of λ and taking as a final estimate the smoothed value at a λ close to 1, in this case λ=0.95 (Boca and Leek, 2018). This method will give a more robust estimate with a reasonable bias.

Here, we present Progressive proportions plot (Prog-Plot; https://github.com/nbellor/progplot), a visual tool to help identify the functional relationship between the co-variate and the null proportion. Prog-Plot was conceived to identify this relationship in the context of the correction proposed by Boca and Leek (2018). In brief, Prog-Plot draws curves of proportions based on different thresholds of λ in a progressive manner, which allows the user to choose the best model fit. Prog-Plot is implemented in the R programming language. Instructions for download and usage can be found at https://github.com/nbellor/progplot.

RESULTS

As a demonstration, we constructed the Prog-Plot for a dataset consisting of 20 paired samples of the GTEx project. Ten of these samples belonged to the Nucleous accumbens tissue of female individuals, and the other 10 to the Putamen tissue. This dataset was previously analyzed by Reyes and Huber (2018), and it is available at https:/doi.org/10.5281/zenodo.1475409.

The pre-processing included a filter of the genes with average counts across all samples of less than 1. This left us with 30,374 genes and 20 samples, as stated previously. These counts are then processed through the DESeq2 pipeline with default settings, and the raw *P*-values are taken to fit the Boca and Leek model (hereafter BL model) correction. Finally, the appropriate calculation is carried out to find the final false discovery rate (FDR)-corrected *P*-value as described later in this section.

Particularly, what we are looking for in the Prog-Plot in Fig. 2 is that the final estimate of the null proportion (the solid line) behavior resembles those of the non-parametric curves (dotted lines). If, for example, the dotted lines exhibited a strong non-linear behavior and the solid line ignored this shape by just estimating a straight line (e.g. with a linear regression), it would be clear that the specified model is inappropriate for the data at hand.

An important assumption of the BL model is that the *P*-values must be independent from the covariate under the null hypothesis. Usually, the histogram of the *P*-values stratified by the covariates is examined in order to evaluate the validity of the hypothesis. Particularly, the histograms must look uniform for larger *P*-values. We plotted the histograms, and they appear to be approximately uniform in Fig. 1.

As suggested by Korthauer et al. (2019), we used the mean gene expression as a covariate, but as the Prog-Plot in Fig. 2 shows, the fit is better with a logarithmic transformation. For this plot, we grouped the response variable based on the quantiles of the covariate given

Statistics Department, Universidad Nacional de Colombia, Ciudad Universitaria, Cra 30 No 45-03, Bogotá 111321, Colombia.

*Author for correspondence (nbellor@unal.edu.co)

N.B., 0000-0002-5089-1484; L.L.-K., 0000-0001-9325-9529

Handling Editor: John Heath
Received 6 June 2022; Accepted 1 December 2022

Journal of Cell Science (2023) 136, jcs260312. doi:10.1242/jcs.260312

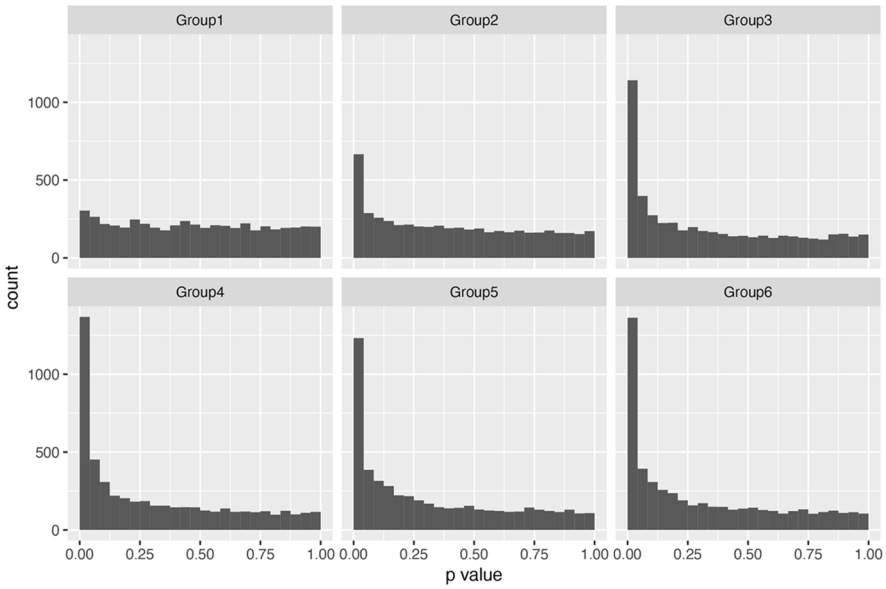

Fig. 1. Histograms of the *P*-values obtained from DESeq2 stratified by the covariate. Diagnostic plot to check the assumption that the *P*-values are independent from the covariate under the null hypothesis. The histograms show counts for *P*-values assigned to each of six groups based on covariate value, with group 1 containing the lowest 17% values of the covariate and group 6 containing the highest 17% of values of the covariate.

that the covariate has a skewed distribution, and if we based the groups on same-length intervals, it would lead to some groups having very few data points to estimate the null proportion.

Please note that in Fig. 2 we are plotting the final estimate $\widehat{\pi_0}$. The final estimate of the q value (the one used to determine whether the gene has differential expression) will be given by multiplying this proportion by the Benjamin and Hochberg (BH) adjusted *P*-value. This means that, for example, if we have a gene with a BH adjusted *P*-value of 0.07 and a logarithm of mean expression of 5 ($\hat{\pi}_0 \approx 0.6$), then its q value will be 0.042. Because of this

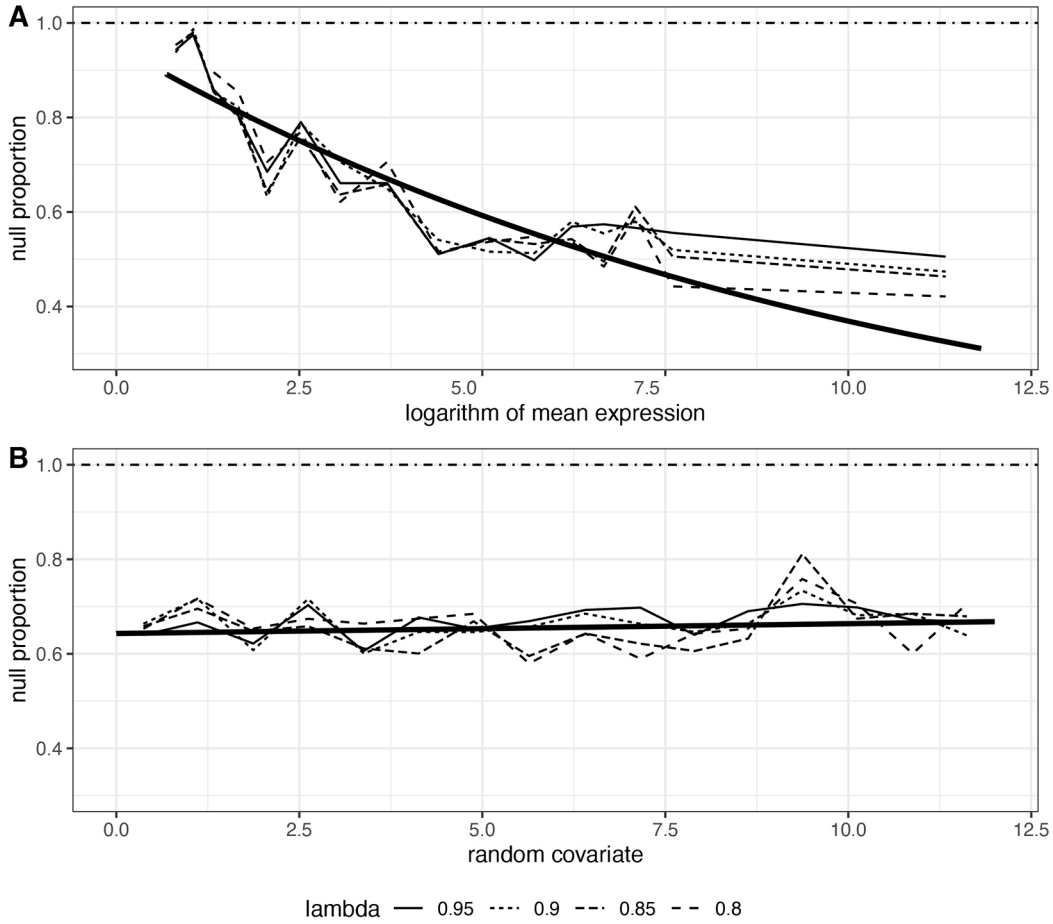

Fig. 2. Prog-Plot. The proportion of estimated *P*-values consistent with the null hypothesis was plotted against the covariate at different threshold λ; the thicker solid black line represents the fit of the BL model. (A) A logistic fit with the logarithm of the mean expression; (B) a logistic model with a randomly generated covariate.

Table 1. Comparison of the number of genes identified as differentially expressed under the BL and BH methods

		BL	
		Not DE	DE
BH	Not DE	27,696	668
	DE	0	2010

Results are shown using the BH and BL *P*-value corrections in the brain samples data set (https:/doi.org/10.5281/zenodo.1475409). The threshold used was 0.05.

relationship, we know that for a pre-specified significance threshold, the number of genes identified as DE will always be more with the BL correction compared to that with the BH correction on its own. For example, in this data set with a significance threshold of 0.05 using the BH correction we found 2010 DE genes, whereas using the BL correction we found 2678 DE genes, as can be seen in Table 1. It is very important to note that this increase in the number of DE genes is not comparable with simply raising the significance threshold for the BH adjusted *P*-values, as the BL correction is mostly adding genes to that list for which we know the test has a higher power (and the BH correction would assume the null proportion to be equal at all levels of the covariate).

DISCUSSION

Prog-Plot (https://github.com/nbellor/progplot) is a visual tool that provides researchers with the capability of specifying an adequate functional relationship between the null proportion and the covariate. Multiple test correction methods are recently incorporating informative covariates through regression models, and this tool will help researchers in taking advantage of the flexibility of these novel methods after verifying in an objective and handy way if the required assumption is met.

MATERIALS AND METHODS

DESeq2

DESeq2 is a widely used technique to identify DE genes proposed by Love et al. (2014). This method takes as input a matrix of raw mapped counts that results from a typical high-throughput sequencing experiment. The first step of this tool is normalizing the counts based on the median of the ratio between the counts themselves and the geometric mean per gene (although other normalizations can also be used). The next step is specifying the generalized linear model (GLM) for these counts, the model for an experiment with two conditions would be the following:

$$\log_2(q_{ij}) = \beta_{i0} + x_i\beta_{i1}, \quad (1)$$

where q_{ij} are the normalized counts of gene i in the sample j, and x_i is an indicator variable of the sample group (usually 0 for control and 1 for treatment). It is worth noting that the counts are assumed to follow a negative binomial distribution.

Apart from the previous model, DESeq2 models the dispersion parameters through another GLM of the gamma family in an iterative manner. The consecutive step consists of the estimation of the coefficients in Eqn 1, which are the log-fold changes (LFCs). This estimation is carried out assuming a normal distribution and the maximum likelihood estimates are used in the prior distribution.

Finally, the statistical tests are carried out where the null hypothesis is that there is no differential expression between the groups (H_0:β_{ir}=0). The test used is based on the Wald statistic, although the implementation allows other tests. A correction must be done owing to the number of multiple tests involved; the default is a BH correction (Benjamini and Hochberg, 1995) with independent filtering (Bourgon et al., 2010). In the following section we discuss a multiple-tests correction that includes a covariate.

Multiple testing correction and plot

Prog-Plot is based on the model proposed by Boca and Leek (2018). The regression model (usually logistic) will be given by the following:

$$\hat{\pi}_0^{\lambda} = \frac{\hat{E}(Y_i|X_i = x_i)}{1-\lambda}, \quad (2)$$

where Y_i=1(P_i>λ) and P_i represents the *P*-value associated to the test performed for the gene i. As Boca and Leek argue, this would only be an estimate for a single λ (Boca and Leek, 2018). When estimating the null proportion through this method, smaller values of λ give biased estimates, but bigger values have an increased variance because of the denominator, hence the values of the null proportion are smoothed over a series of threshold λ. Ultimately, the final estimation of the null proportion will be taken as the smoothed value at λ=0.95.

Given that the response of the model is a binary variable, it is not possible to directly plot the response in any useful way. However, given that we usually work with a considerable amount of hypothesis, we can group the binary response into a proportion for short intervals related to the covariate. These local estimates of the null proportion are specific to a single threshold and form a curve that describes the relationship with the covariate. Additionally, we also plot the proportions curve for increasingly higher thresholds as the final estimate is based on a smooth estimate of these curves.

Acknowledgements

Some of the text and figures in this article form part of the master's thesis of Nicolás Bello at the Universidad Nacional de Colombia, sede Bogotá. We thank the reviewers of Nicolás Bello's master's thesis for their valuable comments. We also thank the reviewers and editors of the journal for reviewing the manuscript, which helped us to greatly improve the quality of the article.

Competing interests

The authors declare no competing or financial interests.

Author contributions

Methodology: N.B.; Software: N.B.; Investigation: N.B.; Writing - original draft: N.B.; Writing - review & editing: N.B., L.L.-K.; Visualization: N.B.; Supervision: L.L.-K.

Funding

This research was supported by master's studentship funding from the Universidad Nacional de Colombia, sede Bogotá (to N.B. and L.L.-K.).

Data availability

All relevant data can be found within the article.

Peer review history

The peer review history is available online at https://journals.biologists.com/jcs/lookup/doi/10.1242/jcs.260312.reviewer-comments.pdf

References

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B Methodol.* **57**, 289-300. doi:10.1111/j.2517-6161.1995.tb02031.x

Boca, S. M. and Leek, J. T. (2018). A direct approach to estimating false discovery rates conditional on covariates. *PeerJ* **6**, e6035. doi:10.7717/peerj.6035

Bourgon, R., Gentleman, R. and Huber, W. (2010). Independent filtering increases detection power for high-throughput experiments. *Proc. Natl Acad. Sci. USA* **107**, 9546-9551. doi:10.1073/pnas.0914005107

Costa-Silva, J., Domingues, D. and Lopes, F. M. (2017). Rna-seq differential expression analysis: an extended review and a software tool. *PLoS ONE* **12**, e0190152. doi:10.1371/journal.pone.0190152

Korthauer, K., Kimes, P. K., Duvallet, C., Reyes, A., Subramanian, A., Teng, M., Shukla, C., Alm, E. J. and Hicks, S. C. (2019). A practical guide to methods controlling false discoveries in computational biology. *Genome Biol.* **20**, 1-21. doi:10.1186/s13059-019-1716-1

Lei, L. and Fithian, W. (2018). Adapt: an interactive procedure for multiple testing with side information. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **80**, 649-679. doi:10.1111/rssb.12274

Love, M. I., Huber, W. and Anders, S. (2014). Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome Biol.* **15**, 550. doi:10.1186/s13059-014-0550-8

Reyes, A. and Huber, W. (2018). Alternative start and termination sites of transcription drive most transcript isoform differences across human tissues. *Nucleic Acids Res.* **46**, 582-592. doi:10.1093/nar/gkx1165

Scott, J. G., Kelly, R. C., Smith, M. A., Zhou, P. and Kass, R. E. (2015). False discovery rate regression: an application to neural synchrony detection in primary visual cortex. *J. Am. Stat. Assoc.* **110**, 459-471. doi:10.1080/01621459.2014.990973