Briefings in Bioinformatics, 20(6), 2019, 2044–2054

doi: 10.1093/bib/bby067
Advance Access Publication Date: 5 July 2019
Review article

Interpretation of differential gene expression results of RNA-seq data: review and integration

Adam McDermaid*, Brandon Monier*, Jing Zhao, Bingqiang Liu and Qin Ma

Corresponding author: Qin Ma, Bioinformatics and Mathematical Biosciences Lab, Department of Agronomy, Horticulture, and Plant Science, South Dakota State University, BioSNTR, Brookings, SD, 57006, USA. Tel.: +1-605-688-6315; E-mail: qin.ma@sdstate.edu
*The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint first authors.

Abstract

Differential gene expression (**DGE**) analysis is one of the most common applications of RNA-sequencing (RNA-seq) data. This process allows for the elucidation of differentially expressed genes across two or more conditions and is widely used in many applications of RNA-seq data analysis. Interpretation of the DGE results can be nonintuitive and time consuming due to the variety of formats based on the tool of choice and the numerous pieces of information provided in these results files. Here we reviewed DGE results analysis from a functional point of view for various visualizations. We also provide an R/Bioconductor package, Visualization of Differential Gene Expression Results using R, which generates information-rich visualizations for the interpretation of DGE results from three widely used tools, *Cuffdiff*, *DESeq2* and *edgeR*. The implemented functions are also tested on five real-world data sets, consisting of one human, one *Malus domestica* and three *Vitis riparia* data sets.

Key words: differential gene expression analysis; differentially expressed genes; bioinformatics tools; visualization and interpretation; *R/Bioconductor package*

Introduction

Next-generation sequencing techniques enable researchers to access far more massive amounts of data than previously available [1–5]. Specifically, RNA-sequencing (RNA-seq) procedures provide an abundance of information regarding the gene expression levels of various organisms across multiple conditions at a high resolution [6–8]. Naturally arising from this information is the concept of (differentially expressed genes) DEGs, which are genes that have expression levels determined to be significantly differentially expressed across two or more conditions [9, 10]. Specific tools have been developed to determine which genes are differentially expressed (Table 1). Differential gene expression (DGE) tools perform statistical tests based on quantifications of expressed genes derived from computational analyses of raw RNA-seq reads (e.g. mapping [10–21] and assembly [10, 22–28]) to determine which genes have a statistically significant difference, while also providing information related to the expression level and pairwise magnitude of difference for each gene. DGE analyses can provide considerable insight into the genetic mechanisms in organisms that are contributing to phenotypic differences, including plant growth patterns [29–31], tumor origin detection [32] and the study of microbiomes [33].

Adam McDermaid is a PhD student in the Department of Mathematics and Statistics at South Dakota State University, Brookings, SD, USA.
Brandon Monier is a PhD student in the Department of Biology and Microbiology at South Dakota State University, SD, USA.
Jing Zhao is an assistant research scientist at Sanford Research and an assistant professor at the Department of Internal Medicine, University of South Dakota Sanford School of Medicine.
Bingqiang Liu is a professor at the School of Mathematics, Shandong University.
Qin Ma is the director of the Bioinformatics and Mathematical Biosciences Lab and an assistant professor at the Department of Agronomy, Horticulture, and Plant Science, South Dakota State University. He is also an adjunct faculty member of the Department of Mathematics and Statistics of SDSU, BioSNTR and Sanford Research, USA.
Submitted: 23 April 2018; **Received (in revised form):** 21 June 2018

Table 1. Citation counts, percentages of commonly referenced DGE tool citations and year of release for edgeR [34], Cuffdiff/Cuffdiff2 [35, 36], DESeq2 [37], limma [38], DEGseq [39], baySeq [40], SAMseq [41], sleuth [42] and NOIseq [43]. All counts were tabulated using the Google Scholar citation counts for the respective tool references as of 2 February 2018

DGE Tool	Citation Count	Percentage	Publish Year
edgeR [34]	7175	32.3%	2010
Cuffdiff/Cuffdiff2 [35, 36]	6103	27.5%	2012/2013
DESeq2 [37]	4355	19.6%	2014
limma [38]	2451	11.0%	2015
DEGseq [39]	1244	5.6%	2009
baySeq [40]	567	2.6%	2010
SAMseq [41]	279	1.3%	2013
sleuth [42]	45	0.2%	2017
NOIseq [43]	39	0.2%	2012

One of the best ways to provide a summary of the DGE results is to generate figures [47, 48], giving a global representation of the expression changes across multiple conditions. DGE tools create output files sharing some information, such as mean gene expression across replicates for each sample, $\log_2$ fold-change (*lfc*) and adjusted *P*-value. However, these output files have many differences in content and structure, which makes generating comprehensive visualizations a time-intensive and potentially challenging task. In this paper, we review common and applicable visualization techniques for DGE results, including descriptions of what information can be interpreted from each figure. The reviewed visualizations are broken down into two tiers based on the information used to generate and the interpretations that can be made using the figure. Tier 1 functions involve more basic visualizations of read count distributions, DEG counts and raw, normalized or transformed read count comparisons. Tier 2 functions require more information and are generated using mean expression values, log fold-changes and adjusted *P*-values.

Additionally, we implement the most useful visualizations into a single R/Bioconductor package, Visualization of Differential Gene Expression Results using R (**ViDGER**), to assist users in generating publication-quality visualizations from *Cuffdiff*, *edgeR* and *DESeq2* capable of providing valuable insight into their generated DGE results (Figure 1). These three selected DGE tools have been shown to be among the highest performing tools for DGE analysis of RNA-seq data [44–46] and contribute to the highest number of citations for DGE tools, representing roughly 80% of all cited DGE tools. The ViDGER package provides six base functionalities for generating information-rich figures derived from the two tiers of reviewed visualization methods. ViDGER also integrates matrix functionalities to provide simultaneous visualization of all pairwise comparisons for three of the base functionalities. In addition to the example data sets provided with the package, ViDGER was extensively tested on five additional data sets from human, *Malus domestica*, and three *Vitis riparia* samples (Example S1).

Visualization methods

Six methods for DGE results visualization will be introduced, and the implementation of each in existing tools is shown in Table 2. While these six functionalities are useful and relatively common, not all are implemented in any of the commonly referenced DGE tools. Most tools have some methods of visualizing the results of DGE analysis. However, none of the tools provides a comprehensive view of using all nine functionalities. Cummerbund [49], a companion tool for Cuffdiff, comes closest to comprehensive visualization, with five of the six reviewed functions. However, this tool is only compatible with Cuffdiff, leaving the other DGE tools with limited capacity for visualizing results. More commonly, a single function is included in the package as a basic method for visualizing the DGE results, as opposed to providing comprehensive visualization of multiple aspects of the DGE results.

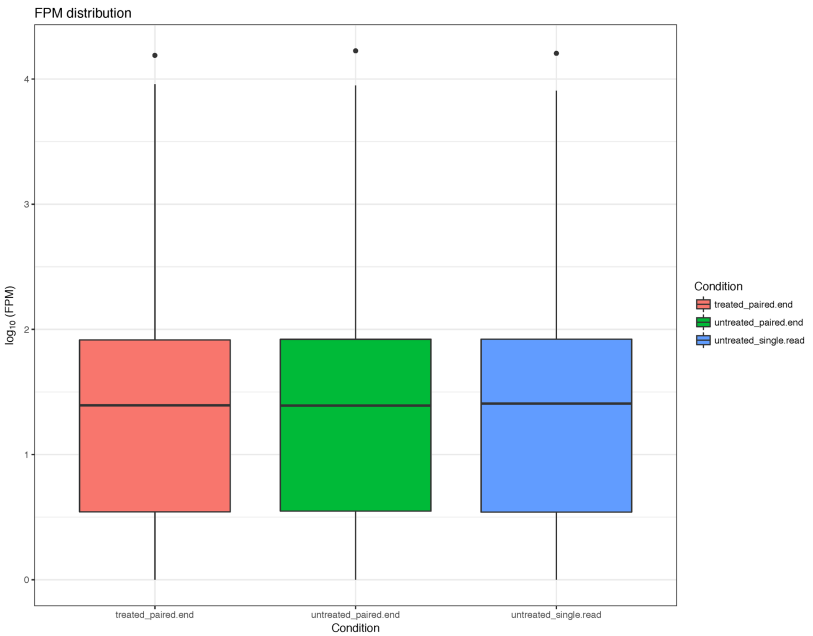

Figure 1. Treatment distributions visualization generated by the ViDGER package using a DESeq2 data set.

Table 2. Nine functions for DGE results analysis and their implementation in existed tools

Function	edgeR	cummeRbund	DESeq2	limma	DEGseq	baySeq	SAMseq	sleuth	NOIseq
Treatment distributions	No	Yes	No	Yes	No	No	No	Yes	Yes
FPKM/CPM scatter plot	No	Yes	No	No	No	No	No	Yes	No
DEG counts	No	Yes	No	No	No	No	No	No	No
MA plot	Yes	Yes	Yes	Yes	No	Yes	No	Yes	Yes
Volcano plot	No	Yes	No	Yes	No	No	No	Yes	No
Four-way plot	No	No	No	No	No	No	No	No	No

Tier I Functions

Tier I consists of more basic functionalities used to visualize raw or normalized expression levels and overall counts of DEGs. These functions do not utilize specific measurements of statistical significance (*P*-value, adjusted *P*-value) or magnitude of the difference (fold-change). Instead, they display expression trends and counts for DEGs. Within this tier, we include methods for (i) visualization of treatment distributions, (ii) comparison of FPKM or CPM and (iii) number of DEGs.

(i) Visualization of treatment distributions

Investigation of the distribution of read counts for each sample can be useful in detecting any abnormalities present in any sample or samples. Ideally, the overall distributions would be similar for all samples displayed. If any sample is drastically different from the others, the user would want to investigate this occurrence further and attempt to rule out any possible biases or erroneous methods that resulted in this difference. To visualize treatment or sample distributions, a few methods can be used. Histograms can provide an appealing way for this purpose, although simultaneously displaying multiple samples or treatment groups can be problematic. The most common implementation of visualizing treatment distributions is through box plots (Figure 1) or their specialized counterparts, such as violin plots or dot plots. While box plots do not directly show an underlying distribution, they can provide the user with information related to the distribution of the quartiles, which can still be useful for this purpose. More useful for this purpose is the modified box plots that show distributions. Violin plots, which are visually and practically similar to box plots, can provide more detailed information about treatment distributions. These figures can use raw reads counts, but more commonly employ some normalization method that controls the range of data points for a more useful and visually appealing graphics. One such normalization method is base-10 logarithm, which is the normalization method used in Figure 1.

(ii) Comparison of expression levels

Another relatively basic visualization method that belongs to Tier 1 is the comparison of expression levels between two samples or two treatment groups. This comparison is generally visualized through the use of scatter plots, where each data point represents a single gene, and its placement indicates its mean respective expression level in two treatments. Scatter plots implemented in this way can be used to compare two treatment groups on a larger scale. Since the axes represent expression levels for their respective category, data points falling along the diagonal would indicate similar expression levels from both groups. Data points above or below the diagonal would mean higher or lower expression levels for the y-axis factor level relative to the x-axis factor level, respectively. When viewing this scatter plot overall, a closer clustering of all data points along the diagonal would indicate two samples or treatment that have highly similar expression patterns across all genes, while more spread of data points from the diagonal would indicate less similar expression levels. To assist in this interpretation, it is common for scatter plots representing expression levels to include a diagonal line for reference. As with the visualization of distributions in section (i), scatter plot comparisons of expression levels frequently use normalized expression values, as opposed to raw counts. This again assists in controlling the range of expression levels to provide a more useful figure. Normalized expression values are often in the form of FPKM (reads per kilobase of transcript per million mapped reads) or CPM (counts per million), and can sometimes even be displayed using a base-10 logarithm scatter plot (Figure 2).

(iii) Number of DEGs

Another useful way to display more general results from DGE analyses is to show the number of DEGs between two treatment groups. Two of the most common are histograms and heatmaps. Histograms can be used to indicate the number of pairwise DEGs for all treatment comparisons simultaneously. This method is highly useful as it directly displays which comparisons are more dissimilar regarding DEGs. Additionally, histograms of this sort can be modified to show the number of upregulated and downregulated DEGs in each comparison. Heatmaps based on the number of DEGs, by comparison, can also be used to display the same information (Figure 3) summarily. Using a spectrum of colors based on the magnitude of the DEG counts, DEG heatmaps can provide a straightforward method that is easily readable and interpretable. For DEG heatmaps, each cell represents the number of DEGs for the respective intersecting row and column. The placement along the chosen color spectrum visually indicates the magnitude, as with Figure 3 where a darker blue indicates a higher number of DEGs and thus more differentially expressed treatment groups. DEG heatmaps do have one distinct downfall related to redundancy. For three factor levels, this figure works well to display the data; however, increasing the number of factor levels results in redundant cells, which are usually left blank as not to mislead users. This method then becomes counterproductive, as it required more effort to interpret the information efficiently. As the number of factor levels grows more substantial, the usefulness of this type of visualization decreases, so it is recommended only for few factor levels.

Tier II Functions

Tier II functions provide more information at the specific gene comparison level. Functions in this tier consider information related to statistical significance, mean expression levels and magnitude of comparison. Consideration of these metrics also allows this tier of functions to provide thresholds based on widely-accepted cutoffs, such as adjusted *P*-values below 0.05

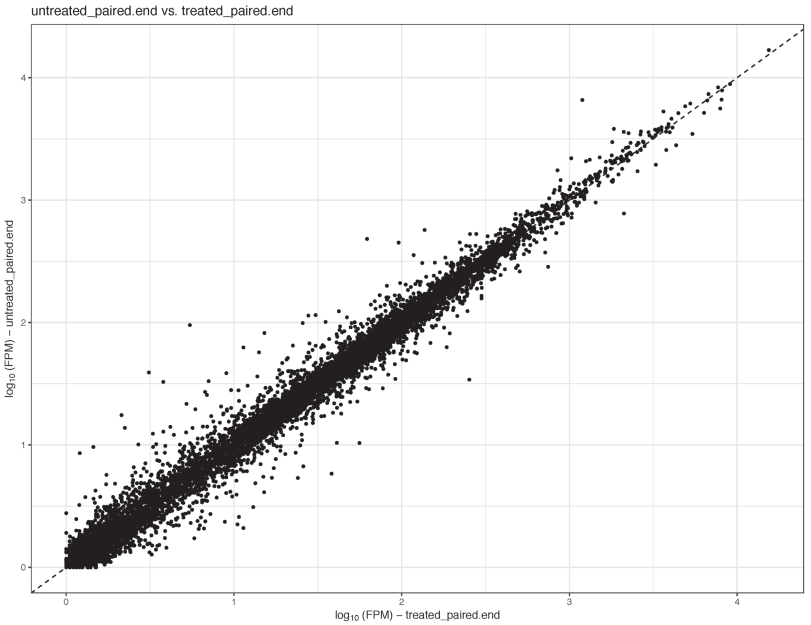

Figure 2. Scatter plot of normalized read counts generated by the ViDGER package using a DESeq2 data set.

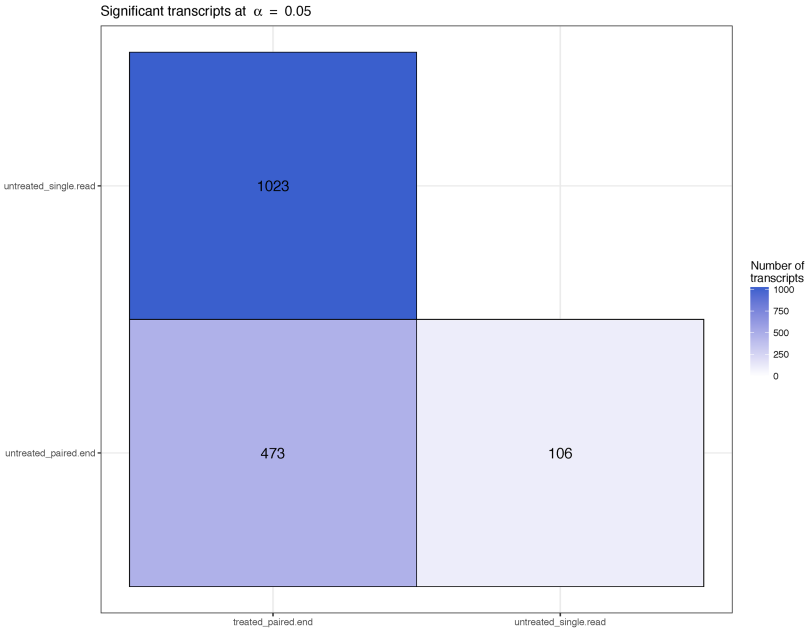

Figure 3. Heatmap of DEG counts by comparison generated by the ViDGER package using a DESeq2 data set with an adjusted P-value cutoff of 0.05 for classification as differentially expressed.

and log fold-changes above 1. The functions in this tier utilized two of these metrics to visualize the results of DGE analysis. Within this tier, we include methods for (iv) fold-change versus normalized mean counts, (v) P-value versus fold-change and (vi) relative comparison of fold-change.

(iv) Fold-change versus normalized mean counts

MA plots are commonly used to represent log fold-change versus mean expression between two treatments (Figure 4). This is visually displayed as a scatter plot with base-2 log fold-change along the y-axis and normalized mean expression along the x-axis. Data points with extreme values along the y-axis represent the genes that have highly differential expression levels (although, not necessarily differentially expressed). Typically, lower mean expression values will have more variability in log fold-change than the higher expression value. This results in a fanning effect of the data points as the graph moves from right to left. Since there are standard thresholds for log fold-changes, MA plots will many times have indications of these cutoffs. However, since this figure does not display any measure of statistical significance, it does not directly indicate which data points are statistically differentially expressed. To accommodate this, some MA plots will color data points to show which have below-threshold adjusted P-values.

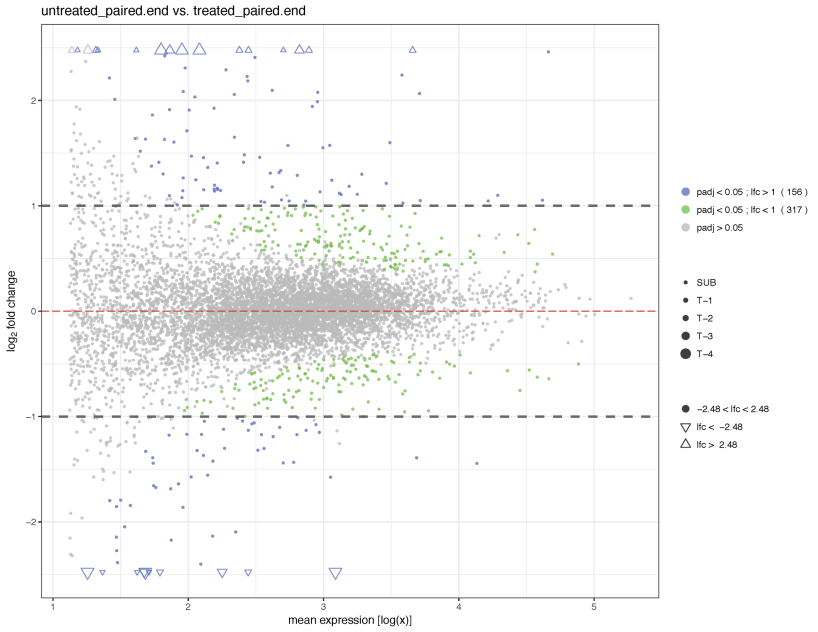

Figure 4. MA plot displaying the log fold-change compared with mean expression generated by the ViDGER package using a DESeq2 data set, with default log fold-change thresholds of −1 and 1.

With each data point again representing a single gene, some valuable information can be extracted from a well-constructed MA plot. A general base-2 log fold-change threshold of 1 indicates which genes either double or halve in the respective comparison. An MA plot with a high number of data points falling above the one threshold on the y-axis would indicate a more significant number of genes being upregulated, while more below −1 would indicate high levels of downregulation in genes. Commonly, MA plots with have a fairly even dispersion relative to the y-axis, which tightens with an increase along the x-axis. Sometimes, biological significance may indicate an expected spread higher or lower on the y-axis than the usual, as may be the case when studying dormant and non-dormant plants. In the scenario where all or most data points fall close to 0 along the y-axis, the two treatment groups would be highly similar in expression patterns.

(v) P-value versus fold-change

Another common comparison of interest between two treatment conditions is the adjusted *P*-value versus log fold-change. This figure is referred to as a volcano plot, as it resembles an exploding volcano, with clusters of data points close to the origin and a fanning effect moving away from this central location (Figure 5). Volcano plots display the statistical significance of the difference relative to the magnitude of difference for every single gene in the comparison, usually through the negative base-10 log and base-2 log fold-change, respectively. Since the *P*-values have a negative transformation, the higher along the y-axis a data point falls, the smaller the *P*-value. It is generally used for volcano plots to include some threshold indicators for adjusted *P*-values to indicate which genes would be considered statistically differentially expressed based on the adjusted *P*-value of their difference between treatments. The log fold-change along the x-axis displays more considerable differences in the extreme values, with data points closer to 0 representing genes that have similar or identical mean expression levels. For volcano plots, a fair amount of dispersion is expected as the name suggests. A wider dispersion indicates two treatment groups that have a higher level of difference regarding gene expression. It is quite rare for a volcano plot to have most, or all data points clustered close to the origin.

(vi) Relative comparison of fold-change

While less common than the other described methods, functionalities that provide a relative comparison of log fold-changes also have broad applicability. A four-way plot is one particular method for visualization of relative fold-change comparisons. In this type of figure, two treatments are compared through their respective log fold-change with a control group (Figure 6). Most commonly, this visualization can be used to compare two distinct treatment groups relative to a control treatment. This comparison is most useful when multiple comparisons are being made against a specific control or corresponding sample. On this type of visualization, the x-axis represents the log fold-change of treatment A with the control, while the y-axis represents the log fold-change of treatment B with the control. Based on log fold-change thresholds, this figure can be broken down into nine distinct regions. The middle region represents genes that have low fold-changes in both treatments relative to the control group. The upper-right and lower-left regions represent the genes that are respectively highly and lowly expressed in both conditions. The upper-left and lower-right regions indicate genes which are highly expressed in one comparison and lowly expressed in the other. The central region on the right and left represent genes with similar expression levels between treatment A and the control group, while treatment B expression levels differ from the control. The central regions on upper and lower areas operate inversely of this. From these regions, a comprehensive view of three-factor levels can be observed. For figures with most or all data points in the central region, both treatments would have similar expressions with the control group. Data points falling along the increasing diagonal from left to right

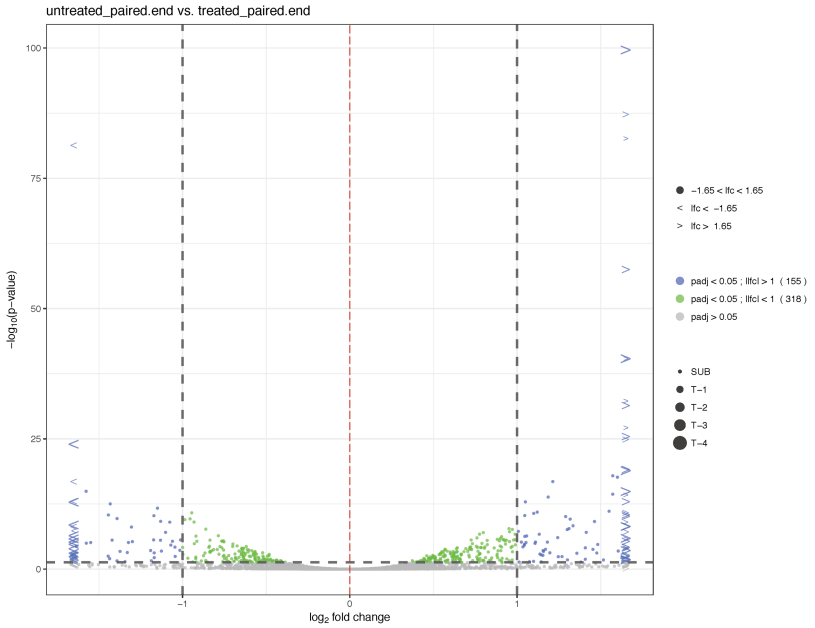

Figure 5. Volcano plot generated by the ViDGER package using a DESeq2 data set, with default log fold-change thresholds of −1 and 1 and an adjusted *P*-value threshold of 0.05.

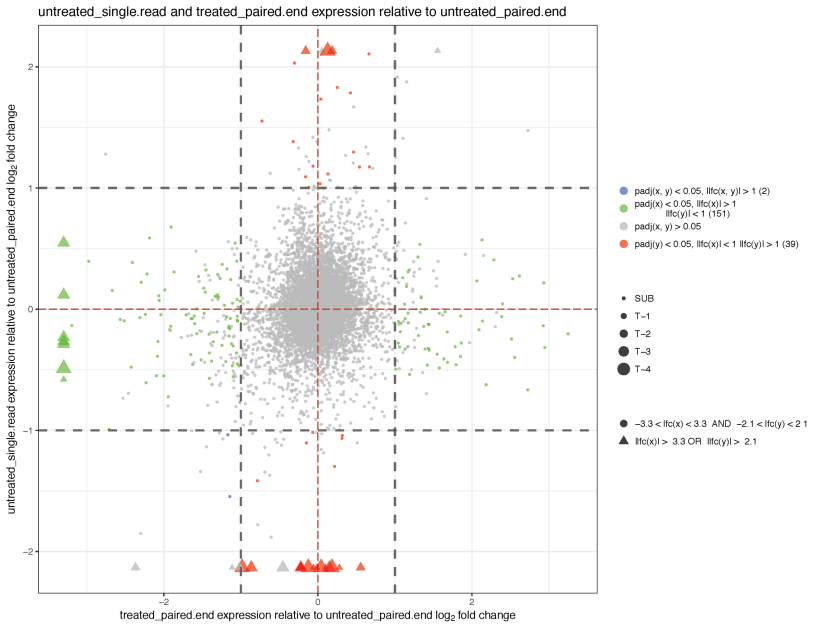

Figure 6. Four-way plot generated by the ViDGER package using a DESeq2 data set, with default log fold-change thresholds of −1 and 1.

would have similarly differing expression levels compared to the control group. Most points falling along the opposite diagonal would represent genes with an inverse relationship relative to the control group.

Integrated visualization package

To assist users in generating the reviewed visualizations for their DGE results, we incorporated the figures into a single R package, ViDGER (Visualization of Differential Gene Expression Results using R). This tool is compatible with DGE results files from the three most widely used DGE tools, *Cuffdiff, edgeR and DESeq2*. ViDGER functions require limited information to generate high-quality visualizations, with the purpose geared towards ease-of-use to quickly generate highly informative visual aids for presentations, posters, and publications (Figure 7). Most ViDGER functions only require user specification of data and data type (i.e. Cuffdiff, DESeq2 or edgeR) and potentially an indication of factor levels of interest.

ViDGER provides visualizations for each of the reviewed visualization methods, including box plots, violin plots, notched box plots and optional dot plot overlays for (i) visualization of treatment distributions, scatter plots for (ii) comparison of expression levels, DEG heatmaps for (iii) visualization of number of DEGs, MA plots for (iv) fold-change versus mean counts, Volcano plots for (v) fold-change versus *P*-value and

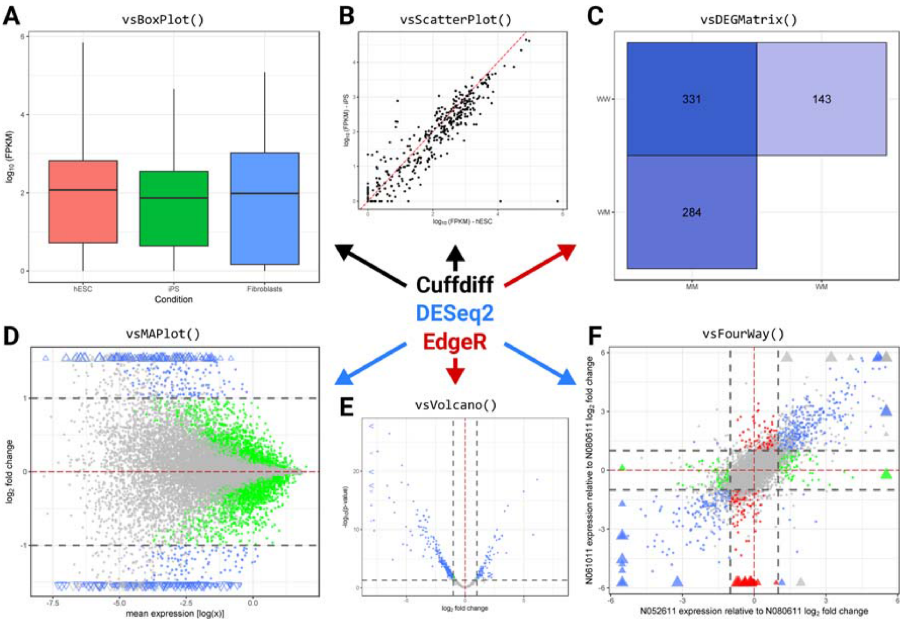

Figure 7. (A) Boxplot generation of RNA-seq data using *vsBoxplot*; (B) scatter plot generation using *vsScatterPlot*; (C) differential gene expression matrix using *vsDEGMatrix*; (D) MA plot generation using *vsMAPlot*; (E) volcano plot generation using *vsVolcano*; (F) four-way plot generation using *vsFourWay*. Arrow and text color refer to visualizations generated using *Cuffdiff* data (black), *DESeq2* data (blue) and *edgeR* data (red).

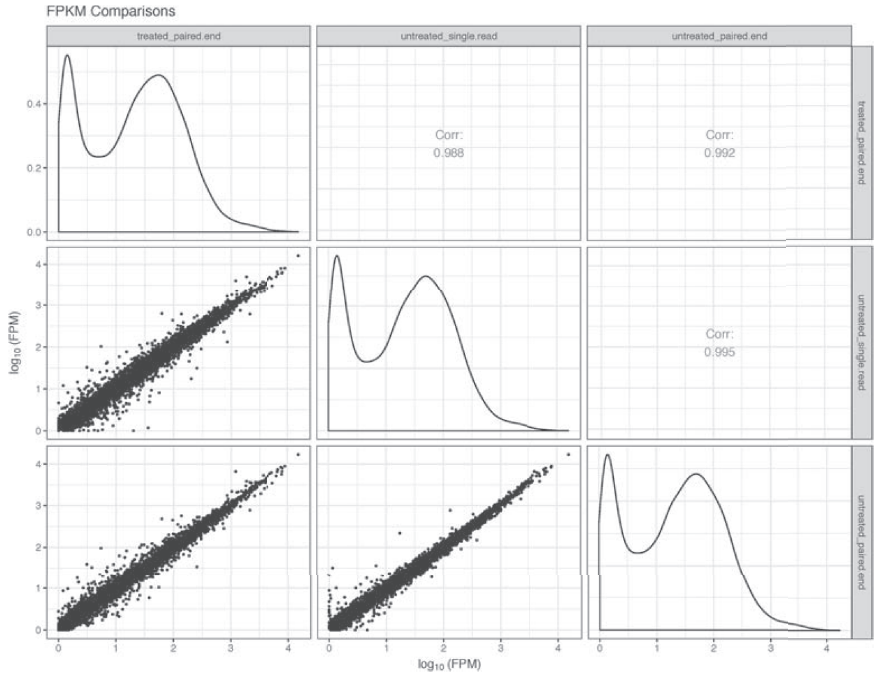

Figure 8. Matrix of all pairwise scatter plots showing normalized expression values generated by the ViDGER package using a DESeq2 data set. In addition to the pairwise scatter plots, density plots are provided along the diagonal and pairwise correlation values are provided in the opposite half of the matrix.

four-way plots for (vi) relative comparison of fold change. For users with specific genes of interest, the scatter plot, MA plot, Volcano plot and four-way plot functionalities allow for a set of user-provided genes to be highlighted in the figure. In addition to the basic functionalities, ViDGER also integrates Scatter plot, MA plot and Volcano plot functionalities into a matrix format displaying all possible pairwise figures in the provided data **(vii–ix)**. The ViDGER package is developed for the R environment (> = 3.5.0) and is freely available through Bioconductor at https://www.bioconductor.org/packages/3.7/bioc/html/vidger.html. More details about the specific ViDGER functions and their application can be found in the *Supplementary Materials*.

(vii) Comparison of FPKM or CPM Treatment combinations

Often, researchers want to visualize multiple pairwise combinations of expression levels at once. While the scatter plot functionality provides an efficient way to compare global

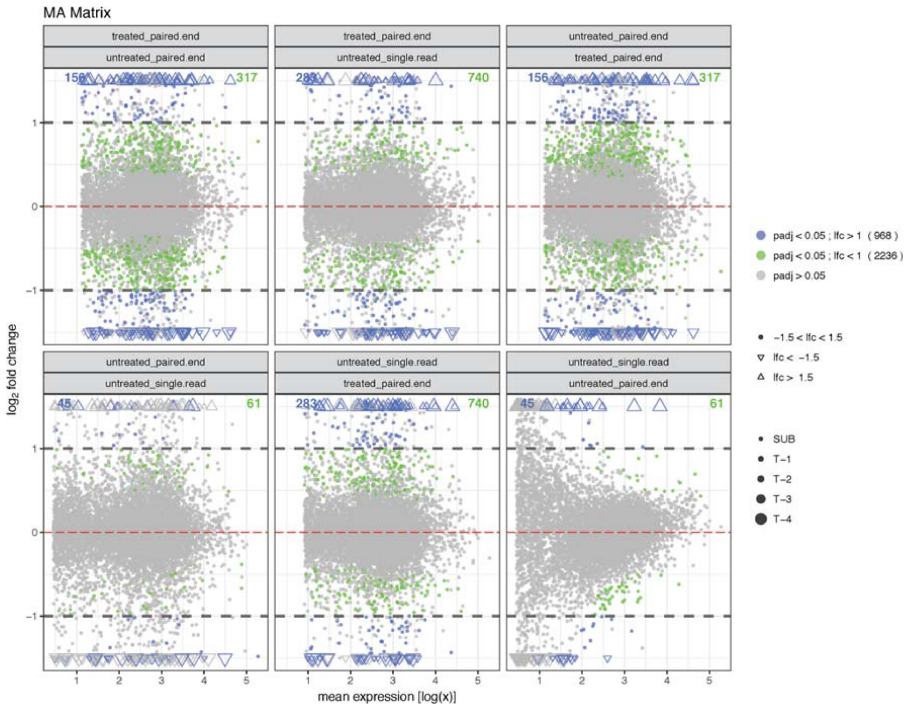

Figure 9. Matrix of all pairwise MA plots showing log fold-change compared with mean expression value generated by the ViDGER package using a DESeq2 data set, with default log fold-change thresholds of −1 and 1.

expression patterns between two specified treatments, its limitation is two treatment comparisons. An efficient way to overcome this hurdle is to generate a matrix of all pairwise comparisons using the scatter plot functionality (Figure 8). This approach integrates the benefits observed through pairwise visualization of expression levels from the scatter plot with the matrix capability of displaying all combinations at once. Each column represents one treatment, each being replicated in a row as well. Each cell represents the pairwise comparison between its row treatment and its column treatment. Since the pairwise matrix of scatter plots has some redundancy from the opposite diagonal display and triviality along the diagonal, these cells are typically replaced with additional information. Commonly, the diagonal cells, which would represent the pairwise comparison of the same treatment, are replaced with expression level density plots. Opposite diagonal cells, which would otherwise represent the same information, are commonly used to display correlation values. These values provide an empirical representation of the overall similarity between the two treatments. A value of 1 indicates identical expression trends, although not necessarily identical expression levels, and a value of −1 indicates perfectly opposite expression trends. Due to the nature of genetic data, the high level of similarity among genetic expressions for the same species will likely result in high correlations. Interpretations of these correlation values are more effectively used to compare the relative similarity between pairwise comparisons. This approach allows for a view of each relative expression pattern and correlation all-in-one visualization.

(viii) Fold-change versus normalized mean counts treatment comparisons

As with the normalized expression scatter plots in (ii), MA plots are only capable of comparing two treatment conditions at once. However, all pairwise comparisons for this figure can be combined into a matrix format to provide all possible combinations simultaneously (Figure 9). For this figure, each cell represents a particular comparison, which is either denoted on a cell-by-cell basis or through the row-column intersection. This visualization enables users to view all pairwise fold-change versus mean expression comparisons at once. Additionally, this method allows for a direct comparison of the pairwise treatment comparisons. Doing so provides an approach to determine which treatment comparisons are more or less similar in both log-fold change and mean expression level. This process, as with the other matrix options, allows users to visualize all their treatment-based comparisons in one figure.

(ix) P-value versus fold-change treatment comparison

Volcano plots encounter the same issues as MA plots in terms of displaying information from only two treatments at once. A similar approach is used to overcome this issue as is used for MA plots: integration of all pairwise comparisons into a single matrix. This matrix functionality enables users to view all pairwise volcano plots simultaneously, giving them a direct look at adjusted *P*-value versus log fold-change for all possible pairwise comparisons (Figure 10). As with the MA plots, each cell of the matrix represents a distinct comparison. Using this functionality, researchers can analyze the pairwise comparisons of *P*-value and fold-change to identify more similar or more different sets of comparisons. Although this option may have limited experience use, it would be useful in situations where users wish to show mass similarity across all comparisons, highlight the individual or limited deviations, or display situations where the comparisons vary widely.

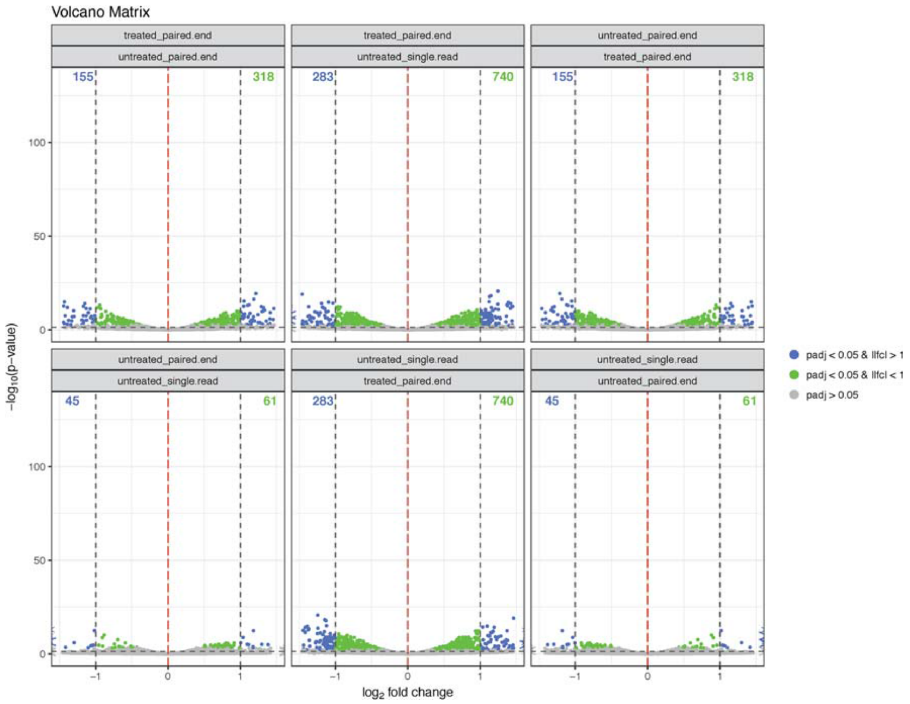

Figure 10. Matrix of all pairwise Volcano plots showing log fold-change versus adjusted *P*-value generated by the ViDGER package using a DESeq2 data set, with default log fold-change thresholds of −1 and 1 and an adjusted *P*-value threshold of 0.05.

Conclusions

DEGs are frequently used to determine genotypical differences between two or more conditions of cells, in support of specific hypothesis-driven studies. Interpretation of this information can benefit significantly from the graphical representation of results files. Tier 1 functions, including those used to visualize reads counts distributions, pairwise expression levels and DEG counts, provide a relatively basic level of information, while Tier 2 functions take additional metrics—such as mean expression levels, fold-changes and *P*-values—to provide more detailed and informative visualizations. Box plots, violin plots, dot plots and read counts histograms can provide insight into the distribution of reads counts for each sample or treatment group. Scatter plots allow users to visualize the overall similarity of expression levels by displaying each gene's expression level in two select treatments or samples. DEG histograms and heatmaps provide a direct representation of the number of DEGs in each comparison. MA and volcano plots are useful in the relative display of mean expression levels, log fold-changes and adjusted *P*-values. Four-way plots, while not applicable for every user, can provide an even higher level of detail by incorporating a third treatment group or sample as a relative or control group.

Although a lot of information and presentation method has been provided in different tools, the integration of these functions in a user-friendly way is still needed. After reviewing six mainstream methods for DEGs result analysis, we have created an R package to assist in the process of generating publication quality figures of DGE results files from *Cuffdiff*, *DESeq2* and *edgeR*. Additionally, we implemented three of the functionalities in matrix form to provide a comprehensive view of all pairwise comparisons. We believe that this package will significantly assist biologists and bioinformaticians in their interpretations of DGE results. Utilizing this package will provide a straightforward method for comprehensively viewing DEGs between samples of interest and allows researchers to generate usable figures for the furthered dissemination of their DGE studies.

Key Points

- **DGE** analysis is one of the most common applications of RNA-seq data. It determines genotypical differences between two or more conditions of cells, in support of specific hypothesis-driven studies.
- The integration and the visualized representation of DGE result analysis functions can facilitate the downstream studies, especially for researchers who have limited computational backgrounds.
- The six reviewed functionalities provide a comprehensive view of DGE results through visualizations.
- The ViDGER R package provides a straightforward method for visualizing DGE results files that from the three most commonly used DGE tools: DESeq2, edgeR and Cuffdiff. Nine functions are provided, including six distinct visualizations with three matrix options.
- The generated visualizations provide comprehensive views of the DGE results files in highly informative, publication-quality figures, all of which can be extracted in multiple formats.

Supplementary Data

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

Acknowledgements

We thank our collaborators for their insightful suggestions on this manuscript and pipeline testing, especially Anne Fennell and Michael Wisniewski for their support in data to extensively test the R package.

Funding

This work was supported by the National Science Foundation/EPSCoR Cooperative Agreement #IIA–1355423 and by the State of South Dakota. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. Support for this project was also provided by an Institutional Development Award (IDeA) from the National Institute of General Medical Sciences of the National Institutes of Health under grant number 5P20GM121341 and Sanford Health–SDSU Collaborative Research Seed Grant Program. This work used the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by National Science Foundation (grant number ACI–1548562). This work is also supported by Hatch Project: SD00H558–15/project accession No. 1008151 from the USDA National Institute of Food and Agriculture. Support for this project was also provided by the National Nature Science Foundation of China (NSFC) [61772313 and 61432010 to B.L.]; and Young Scholars Program of Shandong University (YSPSDU, 2015WLJH19). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health or the U.S. Department of Agriculture.

References

1. Goodwin S, McPherson JD, McCombie WR. Coming of age: ten years of next-generation sequencing technologies. *Nat Rev Genet* 2016;**17**:333–51.
2. Marioni JC, Mason CE, Mane SM, *et al*. RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res* 2008;**18**:1509–17.
3. Miller JA, Menon V, Goldy J, *et al*. Improving reliability and absolute quantification of human brain microarray data by filtering and scaling probes using RNA-Seq. *BMC Genomics* 2014;**15**:154.
4. Nagalakshmi U, Wang Z, Waern K, *et al*. The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science* 2008;**320**:1344–9.
5. Van Dijk EL, Auger H, Jaszczyszyn Y, *et al*. Ten years of next-generation sequencing technology. *Trends Genet* 2014;**30**: 418–26.
6. Finotello F, Di Camillo B. Measuring differential gene expression with RNA-seq: challenges and strategies for data analysis. *Brief Funct Genomics* 2015;**14**:130–42.
7. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics* 2009;**10**: 57–63.
8. Ozsolak F, Milos PM. RNA sequencing: advances, challenges and opportunities. *Nat Rev Genet* 2011;**12**:87–98.
9. Oshlack A, Robinson MD, Young MD. From RNA-seq reads to differential expression results. *Genome Biol* 2010; **11**:220.
10. Pertea M, Kim D, Pertea GM, *et al*. Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. *Nat Protoc* 2016;**11**:1650–67.
11. Li B, Dewey CN. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinform* 2011;**12**:323.
12. Wu J, Anczukow O, Krainer AR, *et al*. OLego: fast and sensitive mapping of spliced mRNA-Seq reads using small seeds. *Nucleic Acids Res* 2013;**41**:5149–63.
13. Bonfert T, Kirner E, Csaba G, *et al*. ContextMap 2: fast and accurate context-based RNA-seq mapping. *BMC Bioinform* 2015;**16**:122.
14. Wang K, Singh D, Zeng Z, *et al*. MapSplice: accurate mapping of RNA-seq reads for splice junction discovery. *Nucleic Acids Res* 2010;**38**:e178.
15. Philippe N, Salson M, Commes T, *et al*. CRAC: an integrated approach to the analysis of RNA-seq reads. *Genome Biol* 2013;**14**:R30.
16. Wu TD, Reeder J, Lawrence M, *et al*. GMAP and GSNAP for genomic sequence alignment: enhancements to speed, accuracy, and functionality. *Methods Mol Biol* 2016;**1418**: 283–334.
17. Dobin A, Davis CA, Schlesinger F, *et al*. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 2013;**29**:15–21.
18. Trapnell C, Pachter L, Salzberg SL. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* 2009;**25**: 1105–11.
19. Kim D, Langmead B, Salzberg SL. HISAT: a fast spliced aligner with low memory requirements. *Nat Methods* 2015;**12**: 357–60.
20. Li B, Ruotti V, Stewart RM, *et al*. RNA-Seq gene expression estimation with read mapping uncertainty. *Bioinformatics* 2009;**26**:493–500.
21. Workbench CG. 'Version 6.5. 1, CLC bio A/S Science Park Aarhus Finlandsgade:10–2.
22. Yuan L, Yu Y, Zhu Y, *et al*. GAAP: genome-organization-framework-assisted assembly pipeline for prokaryotic genomes. *BMC Genomics* 2017;**18**:952.
23. Ye C, Hill CM, Wu S, *et al*. DBG2OLC: efficient assembly of large genomes using long erroneous reads of the third generation sequencing technologies. *Sci Rep* 2016; **6**:31900.
24. Goodwin S, Gurtowski J, Ethe-Sayers S, *et al*. Oxford Nanopore sequencing, hybrid error correction, and de novo assembly of a eukaryotic genome. *Genome Res* 2015;**25**:1750–6.
25. Chang Z, Li G, Liu J, *et al*. Bridger: a new framework for de novo transcriptome assembly using RNA-seq data. *Genome Biol* 2015;**16**:30.
26. Grabherr MG, Haas BJ, Yassour M, *et al*. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol* 2011;**29**:644–52.
27. Pertea M, Pertea GM, Antonescu CM, *et al*. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat Biotechnol* 2015;**33**:290–5.
28. Ji P, Zhang Y, Wang J, *et al*. MetaSort untangles metagenome assembly by reducing microbial community complexity. *Nat Commun* 2017;**8**:14306.
29. Tello-Ruiz MK, Stein J, Wei S, *et al*. Gramene 2016: comparative plant genomics and pathway resources. *Nucleic Acids Res* 2015;**44**:D1133–D1140.
30. Stelpflug SC, Sekhon RS, Vaillancourt B, *et al*. An expanded maize gene expression atlas based on RNA sequencing and its use to explore root development. *Plant Genome* 2016;**9**:1–16.

31. Yang J, Liu D, Wang X, *et al*. The genome sequence of allopolyploid Brassica juncea and analysis of differential homoeolog gene expression influencing selection. *Nat Genetics* 2016;**48**:1225.
32. Tang W, Wan S, Yang Z, *et al*. Tumor origin detection with tissue-specific miRNA and DNA methylation markers. *Bioinformatics* 2017.
33. Niu S-Y, Yang J, McDermaid A, *et al*. Bioinformatics tools for quantitative and functional metagenome and metatranscriptome data analysis in microbes. *Brief Bioinform* 2017;**18**:1–15.
34. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 2010;**26**: 139–40.
35. Trapnell C, Hendrickson DG, Sauvageau M, *et al*. Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nat Biotechnol* 2013;**31**:46.
36. Trapnell C, Roberts A, Goff L, *et al*. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat Protoc* 2012;**7**: 562–78.
37. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* 2014;**15**:550.
38. Ritchie ME, Phipson B, Wu D, *et al*. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res* 2015;**43**:e47–7.
39. Wang L, Feng Z, Wang X, *et al*. DEGseq: an R package for identifying differentially expressed genes from RNA-seq data. *Bioinformatics* 2009;**26**:136–8.
40. Hardcastle TJ, Kelly KA. baySeq: empirical Bayesian methods for identifying differential expression in sequence count data. *BMC Bioinformatics* 2010;**11**:422.
41. 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.
42. Pimentel H, Bray NL, Puente S, *et al*. Differential analysis of RNA-Seq incorporating quantification uncertainty. *Nat Methods* 2017;14:687.
43. Tarazona S, García F, Ferrer A, *et al*. NOIseq: a RNA-seq differential expression method robust for sequencing depth biases. *EMBnet.journal* 2012;**17**:18–9.
44. Sahraeian SME, Mohiyuddin M, Sebra R, *et al*. Gaining comprehensive biological insight into the transcriptome by performing a broad-spectrum RNA-seq analysis. *Nat Commun* 2017;**8**:59.
45. Seyednasrollah F, Laiho A, Elo LL. Comparison of software packages for detecting differential expression in RNA-seq studies. *Brief Bioinform* 2013;**16**:59–70.
46. Kvam VM, Liu P, Si Y. A comparison of statistical methods for detecting differentially expressed genes from RNA-seq data. *Am J Bot* 2012;**99**:248–56.
47. Perkel JM. Data visualization tools drive interactivity and reproducibility in online publishing. *Nature* 2018; **554**:133–4.
48. Tao Y, Liu Y, Friedman C, *et al*. Information visualization techniques in bioinformatics during the postgenomic era. *Drug Discov Today Biosilico* 2004;**2**:237–45.
49. Goff L, Trapnell C, Kelley D. cummeRbund: analysis, exploration, manipulation. In: *and visualization of Cufflinks high-throughput sequencing data*, 2013.