bioRxiv preprint doi: https://doi.org/10.1101/656686; this version posted May 31, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

3D RNA-seq - a powerful and flexible tool for rapid and accurate differential expression and alternative splicing analysis of RNA-seq data for biologists

Wenbin Guo[1], Nikoleta Tzioutziou[1], Gordon Stephen[2], Iain Milne[2], Cristiane Calixto[1], Robbie Waugh[1,3], John W. S. Brown[1,3] and Runxuan Zhang[2]

[1] Division of Plant Sciences, University of Dundee at the James Hutton Institute, Dundee DD2 5DA, UK.
[2] Information and Computational Sciences, The James Hutton Institute, Dundee DD2 5DA, UK.
[3] Cell and Molecular Sciences, The James Hutton Institute, Dundee DD2 5DA, UK.

Correspondence to:

Dr. Runxuan Zhang

E-mail: runxuan.zhang@hutton.ac.uk

Tel No.: +44-1382-568886

Prof John WS Brown

E-mail: j.w.s.brown@dundee.ac.uk

Tel No.: +44-1382-568777

Abstract

RNA-sequencing (RNA-seq) analysis of gene expression and alternative splicing should be routine and robust but is often a bottleneck for biologists because of different and complex analysis programs and reliance on skilled bioinformaticians to perform the analysis. To overcome these issues, we have developed the "*3D RNA-seq*" App, an R shiny App which provides an easy-to-use, flexible and powerful tool for the three-way differential analysis: Differential Expression (DE), Differential Alternative Splicing (DAS) and Differential Transcript Usage (DTU) of RNA-seq data. The full analysis is extremely rapidand can be done within hours. The program integrates *Limma*, a state-of-the-art, highly rated differential expression analysis tool and adopts best practice for RNA-seq analysis. It runs the analysis through a user-friendly graphical interface, can handle complex experimental designs, allows user

bioRxiv preprint doi: https://doi.org/10.1101/656686; this version posted May 31, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

setting of statistical parameters, visualizes the results through graphics and tables, and generates publication quality figures such as heat-maps, expression profiles and GO enrichment plots. The utility of *3D RNA-seq* is illustrated by analysis of Arabidopsis and mouse RNA-seq data. The program is designed to be run by biologists with minimal bioinformatics experience (or by bioinformaticians) allowing lab scientists to take control of the analysis of their RNA-seq data.

Introduction

RNA-seq is generally considered the method of choice to analyse gene expression but is often a source of frustration for experimental biologists. Analysis of RNA-seq data for most biologists is a bottleneck because of reliance on the skills of often over-stretched bioinformaticians who are needed to process large datasets and apply complex analytical programs to experimental data. Many RNA-seq differential analysis programs do not have the flexibility to handle complex experimental designs (such as time-course or developmental series data) and are error prone (Love *et al.*, 2014; Hardcastle and Kelly, 2010; Anders *et al.*, 2012; Nowicka and Robinson, 2016). Results can be inconsistent due to the use of multiple different combinations of tools or pipelines by different bioinformaticians. In addition, despite the ever-increasing appreciation that alternative splicing (AS) is an important level of post-transcriptional regulation, most RNA-seq analyses still focus on the gene expression level thereby losing important information. RNA-seq data, however, contains information that allows the quantification of expression of individual genes and transcripts and the detection of alternative splicing. The availability of programs such as *Salmon* (Patro *et al.*, 2017) and Kallisto (Bray *et al.*, 2016) to quantify transcript and gene level expression accurately and rapidly, allows transcript level analyses to be both feasible and routine.

3D RNA-seq is an interactive web application tool for RNA-seq analysis that is implemented using the R Shiny App. It is developed to carry out a) differential expression (DE) analysis of genes and transcripts; b) differential alternative splicing (DAS) and isoform switch (IS) analysis and c) differential transcript usages (DTU) for RNA-seq data, thus *3D RNA-seq*. The definitions of DE, DAS, DTU and IS are explained in Figure 1. The program integrates the state-of-the-art, highly rated differential expression analysis tool, *Limma* (Ritchie *et al.*, 2015; Law *et al.*, 2014), and adopts current best practice for RNA-seq analysis. It also integrates downstream analysis steps and programs to deliver analyses such as up- and down-regulation of differentially expressed genes, differentially alternatively spliced genes and isoform switch analysis (Figure 2) (Sebestyén *et al.*, 2015; Guo *et al.*, 2017). *3D RNA-seq* can be used regardless of the type of sample or species under investigation, it can handle complex experimental designs and standardizes the analysis process. An easy-to-use graphical interface takes users through the different steps of the analysis, visualizes the intermediate

bioRxiv preprint doi: https://doi.org/10.1101/656686; this version posted May 31, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

and final results through graphics and tables, and generates publication quality figures such as heat-maps and expression profiles. The input to *3D RNA-seq* is a set of transcript quantifications in transcripts per million (TPM) generated by rapid and accurate alignment-free programs, currently *Salmon* (Patro *et al.*, 2017) or *Kallisto* (Bray *et al.*, 2016). The data is pre-processed by generating standardised read counts, it is then adjusted by removing low expressed transcripts and batch effects, followed by data normalisation to reduce technical variances. Then, *3D RNA-seq* sets up statistical models with the user specifying experimental factors, comparisons (contrast groups) and parameters, runs the analysis and generates a full report. In a typical analysis, transcript quantification takes up to two days, and the differential expression analysis and report generation using *3D RNA-seq* takes a few hours (3-Day RNA-seq). There are five main advantages of *3D RNA-seq*: 1) accessibility, ease-of-use and flexibility in setting cut-offs and parameters allows experimental biologists to control the analysis of their own data, 2) acceleration of RNA-seq analyses, 3) ability to handle complex experimental designs, 4) transcript level analysis for accurate differential expression and differential alternative splicing, and 5) the potential to provide an analysis platform to bring consistency in RNA-seq analysis.

Methods

Input files for 3D RNA-seq App

Three types of input files are required for *3D RNA-seq* analysis (Supplementary Figure 1). 1) A meta-data table in “csv” (comma delimited) format containing the information of experimental design, including conditions/treatments, biological replicates, sequencing replicates (if they exist) and quantification file names (Supplementary Figure 1A). 2) Transcript quantification outputs generated using programs *Salmon* (Patro *et al.*, 2017) or Kallisto (Bray *et al.*, 2016) (Supplementary Figure 1B) in conjunction with a reference transcriptome or Reference Transcriptome (Figure 2) which contains a comprehensive set of transcript sequences for the organism under study. *Salmon* and *Kallisto* will generate a quantification file (“quant.sf” from *Salmon* and “abundance.h5”/“abundance.tsv” from *Kallisto*) for each sample, which can be read into the *3D RNA-seq* App and generate read counts and TPMs using the tximport R package (Soneson *et al.*, 2016). 3) A transcript mapping file consisting of a “csv” spreadsheet with the first column listing transcript names and second column listing gene IDs. This relates transcript names to gene IDs in order to summarise transcript level quantifications to gene level expression (Supplementary Figure 1C). Alternatively, this information can be extracted from the “gtf” or “fa” files of the transcriptome and uploaded to *3D RNA-seq* App. For a “gtf” file, transcript-gene mapping will be generated from the “gene_id” and “transcript_id” tags in the last column while for a “fa” file, the information can be generated from the tags in the

bioRxiv preprint doi: https://doi.org/10.1101/656686; this version posted May 31, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

description lines of transcript sequences (Supplementary Figure 1D and E). All the inputs can be easily uploaded to the App by clicking "action" widgets and selecting corresponding files from a local computer.

Data pre-processing with optimal quality control

Once read counts and TPMs are obtained from *Salmon* and *Kallisto*, the data is pre-processed in various steps to reduce noise (e.g. removal of low expressed genes and transcripts) and technical variance (e.g. batch effects). In each step, interactive visualization plots are produced to facilitate the optimization of parameters for pre-processing. 1) Sequencing replicates, if present, will be merged to increase sequencing depth. 2) Unreliable lowly expressed transcripts are removed. Read counts follow approximately the negative binomial distribution. The variance of $\log_2$ transformed read counts decreases monotonically with the increase of mean (Supplementary Figure 2) (Benaroya and Mi Han, 2005; Law *et al.*, 2014; McCarthy *et al.*, 2012). However, the expression of lowly expressed transcripts follows a different distribution. This causes a drop of mean-variance trend towards low values of $\log_2$ read counts (Supplementary Figure 2A). This problem can be solved by removal of lowly expressed transcripts on the basis that an expressed transcript should have a minimum count per million (CPM), *n*, in at least *m* samples. These cut-offs can be adjusted by the user and immediately visualized on mean-variance plots. In this way, the user can quickly define cut-offs to optimise the plot until the decreasing trend at the low expression end of the mean-variance plot is removed (Supplementary Figure 2B). A gene is defined as expressed if any of its transcripts are expressed using the above criteria. 3) Similarities and differences between samples are visualized by Principal component analysis (PCA) (Supplementary Figure 3). PCA is a method to project data variance of thousands of variables (transcripts/genes) to principal component (PC) dimensions from highest to lowest data variability. Therefore, the first few PCs typically reflect the major variance in an experiment. . 4) Batch effects are identified and removed. PCA plots identify the major differences between the samples in an unsupervised manner and effectively highlight whether the RNA-seq data is affected by batch effects, caused by biological replicates being processed, for example, in different laboratory conditions. Compared with random noise, batch effects can be distinguished due to the systematic biases across the biological replicates (Supplementary Figure 3A). When separations of biological replicates are observed in the PCA plot, the batch effect can be corrected by different methods; our preferred method is the RUVSeq R package, which is used to estimate a batch effect term which can be incorporated into the design matrix with the main factors in linear regression models for 3D analysis (Risso *et al.*, 2014) (Supplementary Figure 3B). (5) Data is normalized to unbiased comparisons across samples. Read counts are normalized by the widely used method Trimmed Mean of M-values (TMM), Relative Log

bioRxiv preprint doi: https://doi.org/10.1101/656686; this version posted May 31, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Expression (RLE) or upper-quartile method (Bullard *et al.*, 2010) and $\log_2$ transformed in $\log_2$-CPM) (Law *et al.*, 2014). Read count distributions before and after normalization can be visualized in the plots.

Principle of 3D analysis: identification of DE, DAS and DTU genes and transcripts

RNA-seq experimental designs often involve a single factor or multiple factors that may affect gene expression. *3D RNA-seq* App provides a flexible way to set up contrast groups where users can select any samples or groups of samples of interest for comparison. Thus, it allows analyses both with simple pair-wise comparisons and complex experimental designs such as time-series, developmental series and multiple conditions (Supplementary Figure 4). For each contrast group, different statistics are defined for robust DE/DAS/DTU (3D) predictions (Figure 1). 1) For differential expression, $\log_2$ fold change (L_2FC) is the difference of $\log_2$-CPM values in contrast groups; 2) for differential alternative splicing, Δ percent spliced (ΔPS) is the difference of PS values which are defined as the ratios of transcript average abundances divided by the average gene abundances; and 3) *p*-values of multiple comparisons are adjusted to control the false discovery rate (FDR) (Benjamini and Yekutieli, 2001).

Stringency of the analysis can be modified by changing the cut-off settings. A gene/transcript is identified as DE in a contrast group if L_2FC of expression is greater than or equal to an established cut-off value (e.g. $L_2FC \geq 1$) and with adjusted *p*-value less than a cut-off (e.g. $p<0.01$ or 0.05) (Figure 1). At the AS level, gene expression is compared to transcript expression in the contrast groups (Ritchie *et al.*, 2015). To identify DAS genes, the expression of each transcript is compared to the weighted average expression of all the transcripts for the same gene (weight on transcript expression variance; the average expression can be treated as gene level expression). The *p*-value of each test is converted to gene-wise *p*-value by using the F-test or Simes method (Ritchie *et al.*, 2015). To identify DTU transcripts, each transcript is compared to the weighted average of all the other transcripts of the same gene (Figure 1). A gene is DAS in a contrast group if the adjusted *p*-value is less than an established cut-off value and at least one transcript of the gene has a ΔPS greater than an established cut-off value (e.g. $\Delta PS > 0.1$). A transcript is DTU if the adjusted *p*-value and ΔPS is less than or greater than the selected cut-off values, respectively. The *3D RNA-seq* App thereby identifies significant DE genes and transcripts, DAS genes and DTU transcripts and these results are saved in summary figures (see Figure 3A and B), along with lists of genes and transcripts and tables with testing statistics.

Transcript isoform switch analysis

Transcript isoform switches (ISs) are a prominent type of DAS within a gene where a pair of alternatively spliced isoforms reverse the order of their relative expression abundances in

bioRxiv preprint doi: https://doi.org/10.1101/656686; this version posted May 31, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

response to stimulus (Guo *et al.*, 2017). In the *3D RNA-seq* analysis pipeline, the iso-kTSP method is introduced to study the ISs between pair-wise conditions in the user-defined contrast groups (Sebestyén *et al.*, 2015) while the Time-Series Isoform Switch (TSIS) method is used to identify the ISs between any consecutive time-points in time-series or developmental data (Guo *et al.*, 2017). Significant ISs are determined by: 1) the probability of switch that indicates the frequency of sample replicates reversing their relative abundance at the switches; 2) the sum of average expression differences before and after a switch; and 3) the Benjamini-Hochberg (BH) adjusted *p*-values of these differences (Benjamini and Yekutieli, 2001). In addition, in the TSIS analysis, two further metrics are used to describe ISs: 4) the number of time-points between two switch points and 5) the Pearson correlation of two transcript isoforms (Guo *et al.*, 2017). Customised cut-offs can be applied to these metrics to define significant ISs.

Outputs and visualization

The *3D RNA-seq* App enables users to save results of significant DE/DAS/DTU genes/transcripts, intermediate data of the whole analysis and publication-quality plots to a local folder by clicking “action” buttons at the various steps in the App. Four folders are created to save 3D analysis outputs, “result”, “figure”, “report” and “data”. The gene and transcript expression in read counts and TPMs, testing statistics and analysis results of 3D analysis will be saved as “csv” files (comma delimited) in the “result” folder. All the figures of data pre-processing, 3D analysis and downstream visualization can be saved as “png” and “pdf” formats with user provided width, height and resolution in the “figure” folder. The intermediate variables generated during 3D analysis will be saved as “.RData” R objects in the “data” folder for R users to carry out further analysis if required. In the last step of 3D analysis, a report in three formats, html, pdf and word, will be generated in the “report” folder using *R Markdown* (Baumer and Udwin, 2015). The report includes sections of “Methods”, “Results”, “Supplementary figures”, “Supplementary material” and “References”, in which the contrast groups and parameters selected by the users are recorded. The publication-quality figures and reports provide the information required in publications.

Significant results between contrast groups can be visualized within the App in Venn diagrams, histograms and volcano plots (e.g. the number of up- and down-regulated DE genes and transcripts, isoform switches etc.) (see Figure 3C-E). Heat-maps are used to visualize co-expression clusters of significant 3D genes and transcripts and investigate their expression pattern changes across conditions (see Figure 4A and B). Individual gene and transcript profiles and ΔPS plots give users intuitive visualization of those with significant abundance and AS changes, thereby providing a means of selecting candidate genes and transcripts for experimental validation and future research. The lists of genes and transcripts with significant

bioRxiv preprint doi: https://doi.org/10.1101/656686; this version posted May 31, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

changes in DE/DAS/DTU can be downloaded for gene ontology (GO) enrichment analysis using appropriate web databases or GO analysis programs, from which the significantly enriched annotation results can be re-imported to the *3D RNA-seq* App to generate GO annotation plots (Supplementary Figure 5 and see Figure 4C and D). To generate publication-quality figures, users can set widths, heights, resolution and colours for histograms, plots, heat-maps etc. to be saved. The interactive reports contain full details of the custom configured parameters, methods and results tables and figures from the entire 3D analysis. The interactive report is a detailed document that guarantees the reproducibility of the analysis and results.

3D RNA-seq adopts the best practice and integrates the state-of-art methods for DE and DAS analysis

3D RNA-seq adopts best practice and integrates many of the start-of-art methods for RNA-seq data pre-processing: 1) Tximport to convert TPM values to read counts while taking transcript length and sequencing depth into account (Soneson *et al.*, 2016); 2) PCA plots to visualize sample variances; 3) RUVSeq to estimate batch effects (Risso *et al.*, 2014); 4) a number of read count normalization methods to correct the variances and reduce the false positives for highly abundant transcript outliers (Bullard *et al.*, 2010); and 5) mean-variance trend plots to filter low expressed genes/transcripts to improve fit to statistical models and remove discrepancies due to read count distribution assumptions (Law *et al.*, 2014).

Limma voom (Law *et al.*, 2014) was chosen as the engine for the DE analysis (DE, DAS and DTU) for four reasons. Firstly, from different studies, *limma* is consistently one of the best performing methods for RNA-seq analysis and has a good control of FDR (Pimentel *et al.*, 2017; Rapaport *et al.*, 2013; Tang *et al.*, 2015). Secondly, it allows both the DE and DAS analysis within the same framework. More importantly, the DE analysis and DAS analysis differentiates genes that are implicated in transcriptional regulation (e.g. by transcriptional factors) and alternative splicing regulation (e.g. by splicing factors). Thirdly, *limma* employs a linear model that runs very quickly and compared with methods requiring bootstrapping, the running time and memory required is significantly reduced. Lastly but most importantly, *limma* allows flexible experimental designs, where any pairs or groups of samples can be compared in contrast to most of the current DE tools which offer limited choices for comparisons.

Availability of the 3D RNA-seq App

The *3D RNA-seq* App is available as a docker image and an R package. 1) A public version of the *3D RNA-seq* App is running at the James Hutton Institute and can be viewed at: https://ics.hutton.ac.uk/3drnaseq. The user can upload input files and carry out the entire analysis within a browser. The intermediate data and results of 3D analysis can be

bioRxiv preprint doi: https://doi.org/10.1101/656686; this version posted May 31, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

downloaded during the final step. This web interface allows biologists to use *3D RNA-seq* without installing it. The *3D RNA-seq* manual can be found at the same web address; 2) The R package *ThreeDRNAseq* can be installed locally and users can run the *3D RNA-seq* App by typing a command "run3DApp()" through RStudio on a local PC. This version is suitable for people with a good knowledge of R and who prefer to run the analysis locally. The user manuals for both using ThreeDRNAseq App and command lines of ThreeDRNAseq R package for 3D analysis is provided at:
https://github.com/wyguo/ThreeDRNAseq/tree/master/vignettes/user_manuals

Results

Application of 3D RNA-seq App to RNA-seq analyses of expression/AS in plants

The *3D RNA-seq* App was applied to selected time-points from time-series RNA-seq data of a study on changes in the Arabidopsis transcriptome in response to cold stress (Supplementary Figure 6) (Calixto *et al.*, 2018, 2019). Briefly, 5-week-old Arabidopsis Col-0 plants were grown at 20°C for 24 hours, then the temperature was reduced to 4°C. Samples were harvested every 3 hours for the last day at 20°C, Day 1 at 4°C and Day 4 at 4°C, yielding 26 time-points in total (Supplementary Figure 6A). The data from six of these time-points was extracted to illustrate the utility of *3D RNA-seq*. The six time-points are 3 and 6 h into the dark period from the 20°C Day, Day 1 and Day 4 at 4°C (time points T2 and T3, T10 and T11, and T19 and T20, respectively, referred to here as time-points T2 and T3 from Day 0, Day 1 and Day 4 (Supplementary Figure 6A). The T2 and T3 time-points represent the equivalent time-point in the three treatments and thereby control for time-of-day expression variation. Transcript quantification was generated using *Salmon* (Patro *et al.*, 2017) and AtRTD2-QUASI (Zhang *et al.*, 2017).

In the data pre-processing procedure, we removed the lowly expressed genes/transcripts using the mean-variance trend plots (Supplementary Figure 2). In the original experiment, one biological replicate was harvested in a different year from the other two which gave rise to batch effects. The batch effects were corrected using the RUVSeq method in the App (Supplementary Figure 3). Expression data was normalised across samples to reduce sequencing biases. These pre-processing steps were essential to reduce the FDR and improve the sensitivity and accuracy for the DE and DAS analysis. We also employed stringent filters to control the FDR of multiple testing. These rigorous steps resulted in high confidence predictions of expression changes at both transcript and gene levels.

Using the *3D RNA-seq* analysis pipeline, contrast groups were set up to compare the equivalent time-points before and after cold stress to control for time-of-day variation in

bioRxiv preprint doi: https://doi.org/10.1101/656686; this version posted May 31, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

expression due to photoperiodic and circadian changes (Calixto *et al.*, 2018). For example, T2 in day 1 at 4°C was compared to T2 in day 0 (20°C), T2 in day 4 at 4°C was compared to T2 in day 0 (20°C) and so on (Supplementary Figure 6B). Other contrast groups can also be set up (Supplementary Figure 4). For example, the mean between T2 and T3 in day 1 at 4°C can be compared to the mean of T2 and T3 in day 0 (20°C) etc. or the difference between T2 and T3 in day 1 at 4°C can be compared to the difference between T2 and T3 in day 0 (20°C) etc. (Supplementary Figure 6C). We identified 5,023 DE genes and 2,346 DAS genes. These included 1,875 genes which were regulated only by AS (no significant differential expression at the gene level) and 4,185 DTU transcripts (Figure 3A and B). In addition, 471 of the DAS genes also had significant abundance (DE) changes across the contrast groups (DE+DAS genes). The abundance of significant DTU transcripts can either change significantly (DE+DTU) or non-significantly (DTU-only) (Figure 1). Output histograms illustrate the variation in up- and down-regulation of DE genes and the number of significant isoform switches between the contrast groups (Figure 3C and D). Volcano plots of fold changes in abundance (log_2FC) against significance (FDR) highlight those DE genes and transcripts with the largest and most significant changes in the data (Figure 3E). Heat maps of clustered co-expressed DE genes and DTU transcripts are shown for the contrast groups of the six time-points in Figure 4A and B. Functional annotation of the DE genes showed enrichment in cold-induced physiological and molecular events such as response to various stresses, transcription and altered ribosome production (Figure 4C). Similarly, DAS genes were significantly associated with the spliceosome, RNA splicing and nucleotide/RNA/mRNA binding terms reflecting cold-induced AS of splicing factors (Figure 4D). These results reflect the analysis of the complete time-series (Calixto et al., 2018). Experimental validation of expression, AS and isoform switches by high resolution RT-qPCR and high resolution RT-PCR have been described previously (Calixto et al., 2018, 2019).

The *TSIS* analysis identified 1,688 significant switches between 1,144 pairs of alternative spliced transcript isoforms in DAS genes with stringent cut-offs: probability > 0.5, sum of average abundance differences > 1 TPM and p-value < 0.05 (Figure 3D). These ISs were related to various AS events and approximately half of the isoforms were protein-coding transcripts with different functions according to the transcript annotations in AtRTD2 (Zhang *et al.*, 2017). The switch frequency along the time-series showed that ISs occurred in the contrast groups comparing 4°C and 20°C time-points (Figure 3D). Thus, a large number of genes and transcripts in Arabidopsis are sensitive to temperature reduction and likely contribute to acclimation and survival through AS regulation.

bioRxiv preprint doi: https://doi.org/10.1101/656686; this version posted May 31, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Application to RNA-seq analyses in animals

To illustrate the utility of the *3D RNA-seq* App in analysing multi-factor RNA-seq data from animals, we re-analysed RNA-seq data which studied the effects of dexamethasone treatment on cortical and hypothalamic neural progenitor/stem cells (NPSCs) from male and female mice (Frahm *et al.*, 2018) The RNA-seq data consisted of male and female cortex and hypothalmus cell cultures treated or untreated with dexamethasone, each with three biological replicates (24 samples in total). The study identified differentially expressed genes common and unique to brain region, gender and dexamethasone treatment (Frahm *et al.*, 2018). The same dataset has also been used to demonstrate the improved resolution of differentially expressed genes using *Sleuth* for transcript-level differential analysis and aggregation of transcript level *p*-values to give gene-level *p*-values (Yi *et al.*, 2018). Using this data, we performed two differential expression comparisons between: 1) the *Sleuth*/aggregated *p*-values method (Pimentel *et al.*, 2017; Yi *et al.*, 2018) and *3D RNA-seq* and 2) the results of the differential analysis in Frahm *et al.*, (2018) and those generated by *3D RNA-seq* expression results.

To compare the results from *3D RNA-seq* and *Sleuth* directly, the *Kallisto* transcript quantifications were downloaded (see Data availability) and pre-processed in *3D RNA-seq*. This identified 43,836 expressed transcripts which had at least 3 samples with ≥ 1 CPM and 12,155 expressed genes which had at least one expressed transcript which were then used in the *Sleuth* and *3D RNA-seq* analyses. In addition, the *Sleuth* analysis in Yi *et al.,* (2018) only examined the effects of dexamethasone and did not distinguish brain region or gender effects and therefore the appropriate contrast groups were set up in *3D RNA-seq* to match the analysis in Yi *et al.,* (2018). We then ran the *Sleuth*/aggregated *p*-values pipeline (see Data availability) and *3D RNA-seq* on the 43,836 expressed transcripts. Using *Sleuth* we identified 3,237 DE genes (Figure 5). GO enrichment analysis was performed using Fisher's exact test in topGO R package (Alexa and Rahnenfuhrer, 2019) in conjunction with genome annotation for mouse org.Mm.eg.db (Carlson, 2019). Significantly enriched GO terms in categories of biological process (BP), cellular component (CC) and molecular function (MF) were determined with FDR < 0.05. Significant enrichment terms relevant to response to stress, immune system, inflammation, hormone response, splicing/spliceosome were extracted to illustrate effects of dexamethasone treatment on gene function (Figure 6). The *3D RNA-seq* analysis used the same contrast groups and a BH adjusted *p*-value cut-off of < 0.05 to identify significant changes. Note that to maintain similar parameters for the direct comparison to *Sleuth*, the $\log_2$ fold change and Δ percent spliced cut-offs were not applied in the *3D RNA-seq* analysis. The results of the *3D RNA-seq* pipeline showed a high degree of similarity with the *Sleuth* pipeline in terms of identified DE genes but in addition resolved differentially alternatively spliced genes from DE genes and identified 1,649 genes with significant expression/AS changes. *3D RNA-seq* identified a total of 4,284 genes with differential

bioRxiv preprint doi: https://doi.org/10.1101/656686; this version posted May 31, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

expression and/or differential alternative splicing of which 3,700 and 896 were DE and DAS genes, respectively (with an overlap of 312 gene – DE+DAS genes) (Figure 5). The *Sleuth* pipeline (Yi et *al*., 2018), which does not identify DAS genes, recovered 3,237 DE genes. Of the total (DE and DAS genes) and DE genes identified by *3D RNA-seq*, 2,635 and 2,346, respectively, were common to both analyses such that 81.4% and 72.5% of the total and DE genes identified by *Sleuth* were identified by *3D RNA-seq* (Figure 5). The *3D RNA-seq* pipeline also identified 5,573 DE transcripts and 1,480 DTU transcripts with adjusted *p*-value < 0.05. Interestingly, 531 of the DAS genes were among the DE genes defined by *Sleuth* – of these, 242 were DE+DAS. Of the 602 DE genes unique to the *Sleuth* analysis, 372 had significant DE/DTU transcripts in *3D RNA-seq* but did not carry over to significant DE or DAS genes (Figure 5).

The resolution of DE and DAS genes by *3D RNA-seq* was also illustrated in the GO enrichment annotations. The DE genes from *Sleuth* had significant terms that related to response to stress, response to hormone and immune system as well as multiple splicing/spliceosome terms indicating alternative splicing regulation of the gene level predictions (Figure 6A). The separation of DE and DAS genes by *3D RNA-seq* resolved the majority of stress, hormone and immune response terms to DE genes (transcriptional regulation) (Figure 6B) and the splicing enrichment terms to the DAS genes (AS regulation) (Figure 6C). In addition, the 1,284 DE genes unique to the *3D RNA-seq* analysis (Figure 5) were enriched for steroid hormone receptor activity (Figure 6D) reflecting the nature of the chemical.

The second comparison exploited the flexibility of *3D RNA-seq* to analyse the effects of multiple factors and detect sex-specific and brain region-specific DE and DAS genes and transcripts in the mouse data. In the original study, the effects of dexamethasone treatment on differential gene expression in NPSCs from different brain (cortical and hypothalamic) and sexes (male and female), were examined (Frahm *et al.*, 2018). We therefore set up contrast groups (i.e. Female.Cortex.Dex vs Female.Cortex.Vehicle (untreated control), Male.Cortex.Dex vs Male.Cortex.Vehicle, Female.Hypothalmus.Dex vs Female.Hypothalmus.Vehicle and Male.Hypothalmus.Dex vs Male.Hypothalmus.Vehicle). The cut-offs were set as adjusted *p-value* < 0.05, L_2FC ≥1 and *ΔPS*≥0.1. Across the four contrast groups, the 3D pipeline identified 930 DE genes and 509 DAS genes, of which 462 were only regulated by AS, and 2,121 DE transcripts and 628 DTU transcripts, of which 455 were AS regulated only. The results of individual contrast groups revealed that 1) both transcription and AS regulation were much more affected in cortex cells than in hypothalamic cells (Figure 7A-C) and 2) more genes and transcripts were down-regulated in cortex cells while in hypothalamic cells up-regulation dominated the expression changes (Figure 7A-C). The relative differences between brain regions at the DE gene level was described by Frahm

bioRxiv preprint doi: https://doi.org/10.1101/656686; this version posted May 31, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

et al., (2018). We also compared the DE genes from the 3D pipeline to those of Frahm *et al.,* (2018) which used a *p-value* cut-off of < 0.05 (note: *p*-values were not adjusted and no L_2FC cut-off was applied). There were 449 DE genes in common and 908 significant DE genes in Frahm *et al.,* (2018) were filtered out in the 3D pipeline due to low L_2FC or insignificant adjusted *p*-values (Figure 7D). Although stringent filters were used, the 3D analysis identified 481 unique DE genes. Finally, the isoform switch analysis identified 63 significant ISs of DAS genes in the contrast groups with switch probability ≥ 0.5, sum of average TPM differences ≥ 1 TPM and BH adjusted *p-value* < 0.05 (Figure 7E). Thus, compared to the Sleuth pipeline, *3D RNA-seq* provides both transcript- and gene-level analysis to resolve both differentially expressed genes and transcripts, and provides information on AS regulation by identifying novel DAS genes, DTU transcripts and ISs. The re-analysis of the mouse data with *3D RNA-seq* demonstrates how novel information can be unlocked from publicly available RNA-seq data which has only been used to analyse differential gene-level expression.

Discussion

3D RNA-seq is easy to use and designed for the maximum take-up by biologists.

The *3D RNA-seq* web-based application requires no installation and no programming skills from users. The whole analysis can be accomplished entirely within a web browser. The pipeline consists of 21 single steps within six tabbed panels. It integrates widgets that provide users with technical summaries of the research behind each step so that they can set the parameters appropriate for their studies. Recommended parameters are also given in the manual and are used as defaults. More importantly, *3D RNA-seq* integrates interactive visualization at every step of the analysis so that users can visualize the changes caused by modification of the parameters. Interactive visualization not only helps the user to understand the method behind this step better, but it also enables users to explore the optimum settings for the analysis, and ultimately offers reassurance and robustness in the results.

3D RNA-seq also allows the analysis of RNA-seq data with simple and complex experimental designs, such as time-series and developmental series data etc. The comparisons (contrast groups) can be set up between any pairs or groups of samples to cater for different investigations, a feature that is important for the universal applicability of *3D RNA-seq* for RNA-seq analysis. The *3D RNA-seq* App can be used to analyse RNA-seq data from any species for which transcript quantifications using *Salmon*/*Kallisto* can be generated. The accuracy of the analysis will depend on the quality and comprehensiveness of the reference transcriptome used to generate the transcript quantifications (Brown *et al.*, 2017; Zhang *et al.*, 2015, 2017) Multiple factors such as missing transcripts, transcript redundancy, transcript

bioRxiv preprint doi: https://doi.org/10.1101/656686; this version posted May 31, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

fragments and variation at the 5' and 3' ends of isoforms can drastically affect accuracy of transcript quantification (Alamancos *et al.*, 2015; Zhang *et al.*, 2015, 2017; Soneson *et al.*, 2019). For many species, comprehensive, optimised transcriptomes do not exist and using partial, incomplete transcriptomes and/or which have not been filtered to remove such erroneous isoform information will also produce inaccurate transcript and AS quantification. As transcriptomes improve for many species, transcript quantification and results from *3D RNA-seq* will also improve.

The speed and accuracy of the *3D RNA-seq* analysis has the potential to revolutionise gene expression research programmes. Generation of transcript quantifications and running *3D RNA-seq* is performed in less than three days. It reduces an analytical step which previously required skilled bioinformaticians and often took many weeks or even months to complete. This is significant not only in terms of satisfaction in result generation for the individual scientists/biologists but also for strategic planning of the research programme where now multiple, consecutive RNA-seq experiments can be planned within the period of funding proposal. More importantly, the rapid turn-over of the analysis provided by *3D RNA-seq* creates a level playing field for research groups with very different access to bioinformatics expertise or resources thereby supporting, in particular, smaller groups and early career scientists.

3D RNA-seq facilitates speedy publication of RNA-seq analysis and improves the transparency and reproducibility of the analysis.

The fast turn-over of the analysis provided by 3D RNASeq will present a significant advantage for speedy publication of the results. In *3D RNA-seq*, multiple types of figures/summaries are generated and saved including commonly used visualizations such as heatmaps, GO enrichment plots, expression profile plots, volcano plots etc. Users can easily generate and save new figures for each new set of parameters. The format, resolution and size of the figure can be customised and previewed in *3D RNA-seq* before saving. In addition, the significant DE/DAS/DTU/IS gene and transcript lists are saved for further interpretation. The final technical report generated in the final step of *3D RNA-seq* records each step in the analysis with the parameters used and integrates all the saved figures. The report is comprehensive, accurate and reproducible including all the information required for "Material and Methods" sections for the RNA-seq analysis as well as figures for the Results and Supplementary Materials. Furthermore, some publications have very poor technical descriptions of the methods and parameters used for RNA-seq analyses. In future, submission of such a report along with manuscripts to scientific journals would facilitate transparency and allow reviewers to identify issues with the analysis and how it relates to other published work.

bioRxiv preprint doi: https://doi.org/10.1101/656686; this version posted May 31, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

3D RNA-seq provides enhanced alternative splicing analysis at the transcript level

Over the last 20 years, genome-wide expression analyses have relied on microarrays and more recently, RNA-seq analyses mostly at the gene level. RNA-seq allows gene expression to be analysed at the level of both genes and transcripts which in turn provides a means to study post-transcriptional processes such as alternative splicing. AS regulation plays key roles in gene expression and novel genes with altered gene expression at AS level have been found in most of the RNA-seq studies. Despite the importance of AS and the potential to include AS analysis routinely in RNA-seq, AS is still largely ignored. For example, 4,065 publications from the Web of Science Core Collection (2008-2019) (https://wok.mimas.ac.uk/) were retrieved with the term "RNA-seq differential gene expression", but only ca. 289 included "differential alternative splicing". *3D RNA-seq* provides by far the most detailed differential expression analysis on the transcript and alternative splicing level. In particular, it allows the identification of both DAS genes and DTU transcripts which are differentiated from DE genes/transcripts and a range of measurements, visualizations and statistical tests provide a comprehensive analysis of alternative splicing alongside differential expression. Additionally, *3D RNA-seq* has also integrated methods that detect isoform switches which can play pivotal roles in re-programming of gene expression through switching of functionally different transcript isoforms between, for example, normal and tumor tissues to provide signatures for cancer diagnostics and prognostics (Sebestyén *et al.*, 2015). With the enhanced AS analysis in *3D RNA-seq*, key and novel genes under AS regulation which underpin important biological processes can be identified and provide new targets for medical intervention or crop improvement.

3D RNA-seq unlocks the discovery potential for RNA-seq data

Finally, the speed of *3D RNA-seq* now makes it feasible for biologists to re-analyse existing or publicly available RNA-seq data to give improved differential expression analysis and novel AS information as demonstrated here for the mouse data. Datasets from different labs can now be quickly compared using the same parameters and without performing new experiments. In addition, with new transcriptome releases, it is feasible to re-analyse datasets and update results. *3D RNA-seq* provides a reproducible platform to standardise analyses and utilise new and existing data for improved resolution and interpretation.

Acknowledgements

This work was supported by funding from the Biotechnology and Biological Sciences Research Council (BBSRC) [BB/P009751/1] to JB; BB/R014582/1 to RW and RZ; BB/S004610/1 (16 ERA-CAPS BARN) to RW; the Scottish Government Rural and Environment Science and Analytical Services division (RESAS) [to RZ, RW and JB]. We thank Craig Simpson, Vivek

bioRxiv preprint doi: https://doi.org/10.1101/656686; this version posted May 31, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Raxwal (Masaryk University, Czech Republic), Pingtao Ding (The Sainsbury Laboratory, Norwich) for testing the *3D RNA-seq* App.

Data availability

The *3D RNA-seq* web interface is available at https://ics.hutton.ac.uk/3drnaseq. The R package version (ThreeDRNAseq) is available on Github at
https://github.com/wyguo/ThreeDRNAseq
Manuals for both versions can be accessed from
https://github.com/wyguo/ThreeDRNAseq/tree/master/vignettes/user_manuals

The *Kallisto* transcript quantifications from the dexamethasone treatment on mice were downloaded from:
https://figshare.com/articles/kallisto_quantifications_of_Frahm_et_al_2017/6203012.
The Sleuth/aggregated *p*-values pipeline is at:
https://pachterlab.github.io/sleuth_walkthroughs/pval_agg/analysis.html.

References

Alamancos,G.P., Pagès,A., Trincado,J.L., Bellora,N., and Eyras,E. (2015) Leveraging transcript quantification for fast computation of alternative splicing profiles. *Rna*, **21**, 1521–1531.

Alexa,A. and Rahnenfuhrer,J. (2019) topGO: enrichment analysis for Gene Ontology. R Packag. version 2.26.0. *R Packag. version 2.36.0*.

Anders,S., Reyes,A., and Huber,W. (2012) Detecting differential usage of exons from RNA-seq data. *Genome Res.*, **22**, 2008–2017.

Baumer,B. and Udwin,D. (2015) R Markdown. *Wiley Interdiscip. Rev. Comput. Stat.*

Benaroya,H. and Mi Han,S. (2005) Probability Models in Engineering and Science. *Mech. Eng.*, **48**, 740.

Benjamini,Y. and Yekutieli,D. (2001) The control of the false discovery rate in multiple testing under dependency. *Ann. Stat.*, **29**, 1165–1188.

Bray,N.L., Pimentel,H., Melsted,P., and Pachter,L. (2016) Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.*, **34**, 525–527.

bioRxiv preprint doi: https://doi.org/10.1101/656686; this version posted May 31, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Brown,J.W.S., Calixto,C.P.G., and Zhang,R. (2017) High-quality reference transcript datasets hold the key to transcript-specific RNA-sequencing analysis in plants. *New Phytol.*

Bullard,J.H., Purdom,E., Hansen,K.D., and Dudoit,S. (2010) Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinformatics*, **11**, 94.

Calixto,C.P.G., Tzioutziou,N.A., James,A.B., Hornyik,C., Guo,W., Zhang,R., Nimmo,H.G., and Brown,J.W.S. (2019) Cold-Dependent Expression and Alternative Splicing of Arabidopsis Long Non-coding RNAs. *Front. Plant Sci.*, **10**, 235.

Calixto,C.P.G., Guo,W., James,A.B., Tzioutziou,N.A., Entizne,J.C., Panter,P.E., Knight,H., Nimmo,H.G., Zhang,R., and Brown,J.W.S. (2018) Rapid and Dynamic Alternative Splicing Impacts the Arabidopsis Cold Response Transcriptome. *Plant Cell*, **30**, 1424–1444.

Carlson,M. (2019) org.Mm.eg.db: Genome wide annotation for Mouse. R package version 3.4.0. *Bioconductor*.

Frahm,K.A., Waldman,J.K., Luthra,S., Rudine,A.C., Monaghan-Nichols,A.P., Chandran,U.R., and DeFranco,D.B. (2018) A comparison of the sexually dimorphic dexamethasone transcriptome in mouse cerebral cortical and hypothalamic embryonic neural stem cells. *Mol. Cell. Endocrinol.*, **471**, 42–50.

Guo,W., Calixto,C.P.G., Brown,J.W.S., and Zhang,R. (2017) TSIS: An R package to infer alternative splicing isoform switches for time-series data. *Bioinformatics*, **33**, 3308–3310.

Hardcastle,T.J. and Kelly,K.A. (2010) baySeq: Empirical Bayesian methods for identifying differential expression in sequence count data. *BMC Bioinformatics*, **11**, 422.

Law,C.W., Chen,Y., Shi,W., and Smyth,G.K. (2014) voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol*, **15**, R29.

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.

McCarthy,D.J., Chen,Y., and Smyth,G.K. (2012) Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids*

bioRxiv preprint doi: https://doi.org/10.1101/656686; this version posted May 31, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Res.

Nowicka,M. and Robinson,M.D. (2016) DRIMSeq: a Dirichlet-multinomial framework for multivariate count outcomes in genomics. *F1000Research*, **5**, 1356.

Patro,R., Duggal,G., Love,M.I., Irizarry,R.A., and Kingsford,C. (2017) Salmon provides fast and bias-aware quantification of transcript expression. *Nat. Methods*, **14**, 417–419.

Pimentel,H., Bray,N.L., Puente,S., Melsted,P., and Pachter,L. (2017) Differential analysis of RNA-seq incorporating quantification uncertainty. *Nat. Methods*, **14**, 687–690.

Rapaport,F., Khanin,R., Liang,Y., Pirun,M., Krek,A., Zumbo,P., Mason,C.E., Socci,N.D., and Betel,D. (2013) Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data. *Genome Biol.*, **14**, R95.

Risso,D., Ngai,J., Speed,T.P., and Dudoit,S. (2014) Normalization of RNA-seq data using factor analysis of control genes or samples. *Nat. Biotechnol.*, **32**, 896–902.

Ritchie,M.E., Phipson,B., Wu,D., Hu,Y., Law,C.W., Shi,W., and Smyth,G.K. (2015) limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res*, **43**, e47.

Sebestyén,E., Zawisza,M., and Eyras,E. (2015) Detection of recurrent alternative splicing switches in tumor samples reveals novel signatures of cancer. *Nucleic Acids Res.*, **43**, 1345–1356.

Soneson,C., Love,M.I., Patro,R., Hussain,S., Malhotra,D., and Robinson,M.D. (2019) A junction coverage compatibility score to quantify the reliability of transcript abundance estimates and annotation catalogs. *Life Sci. Alliance*.

Soneson,C., Love,M.I., and Robinson,M.D. (2016) Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. *F1000Research*, **4**, 1521.

Tang,M., Sun,J., Shimizu,K., and Kadota,K. (2015) Evaluation of methods for differential expression analysis on multi-group RNA-seq count data. *BMC Bioinformatics*, **16**, 360.

Yi,L., Pimentel,H., Bray,N.L., and Pachter,L. (2018) Gene-level differential analysis at transcript-level resolution. *Genome Biol.*, **19**, 53.

Zhang,R., Calixto,C.P.G., Marquez,Y., Venhuizen,P., Tzioutziou,N.A., Guo,W., Spensley,M., Entizne,J.C., Lewandowska,D., Have,S. Ten, Frey,N.F., Hirt,H., James,A.B.,

bioRxiv preprint doi: https://doi.org/10.1101/656686; this version posted May 31, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Nimmo,H.G., Barta,A., Kalyna,M., and Brown,J.W.S. (2017) A high quality Arabidopsis transcriptome for accurate transcript-level analysis of alternative splicing. *Nucleic Acids Res.*, **45**, 5061–5073.

Zhang,R., Calixto,C.P.G., Tzioutziou,N.A., James,A.B., Simpson,C.G., Guo,W., Marquez,Y., Kalyna,M., Patro,R., Eyras,E., Barta,A., Nimmo,H.G., and Brown,J.W.S. (2015) AtRTD - a comprehensive reference transcript dataset resource for accurate quantification of transcript-specific expression in Arabidopsis thaliana. *New Phytol.*, **208**, 96–101.

bioRxiv preprint doi: https://doi.org/10.1101/656686; this version posted May 31, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Figures

DE genes/transcripts
Gene/transcript-based expression change between conditions
⇒ gene/transcript based p-value

DAS genes
Compare each transcript change to gene change
⇒ *n* p-values of *n* transcripts vs gene
⇒ Summarise to one gene-level p-value with F-test or Simes method

DTU transcripts
Compare each transcript change to the average change of all the remaining transcripts
⇒ *n* p-values of *n* transcripts

Isoform switches (ISs)
A pair of alternatively spliced isoforms reverse the order of their relative abundance

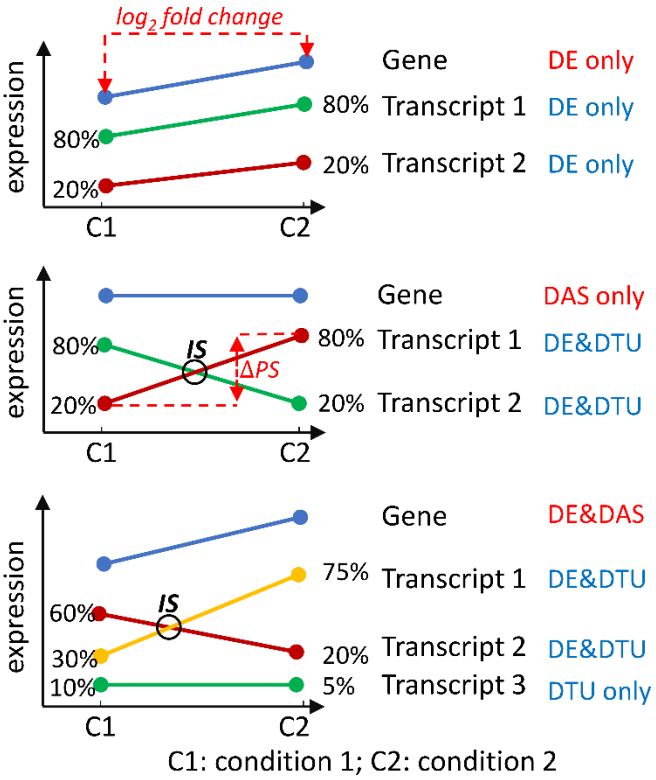


Figure 1. Definitions and criteria for identification of significant DE and DAS genes and DE and DTU transcripts. A) DE genes and transcripts are those whose abundance between conditions changes, as measured by changes in $\log_2$ fold change (L_2FC). For significantly DE genes and transcripts, the default setting is $L_2FC \geq 1$ with an adjusted p-value of ≤ 0.01 or ≤ 0.05. B) DAS genes must have more than one transcript and are determined by comparing the expression changes between individual transcripts to the gene level between conditions. The change in percent spliced (ΔPS) is calculated and for a gene to be DAS, the default is that at least one transcript must have a ΔPS ≥ 0.1 with a pre-set p-value cut-off. C) DTU transcripts are those transcripts which show different expression behaviour from the other transcripts of the gene. They are determined by comparing the change in expression of each transcript to the average expression change of all of the remaining transcripts of the gene. With these criteria, A) DE only genes are those where the gene and transcript expression levels change significantly but to the same degree such that transcripts do not differ from one another and are DE only. B) DAS only genes are those where the gene expression level does not change significantly but that of at least one transcript does. C) DE+DAS genes show both gene level expression changes and different changes of at least one transcript. A-C) The abundance of a DTU transcript can change significantly (DE+DTU) or may not change significantly (DTU only). Isoform switches (ISs) happen when two pair of transcripts reverse their relative abundance across different conditions or time points

bioRxiv preprint doi: https://doi.org/10.1101/656686; this version posted May 31, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

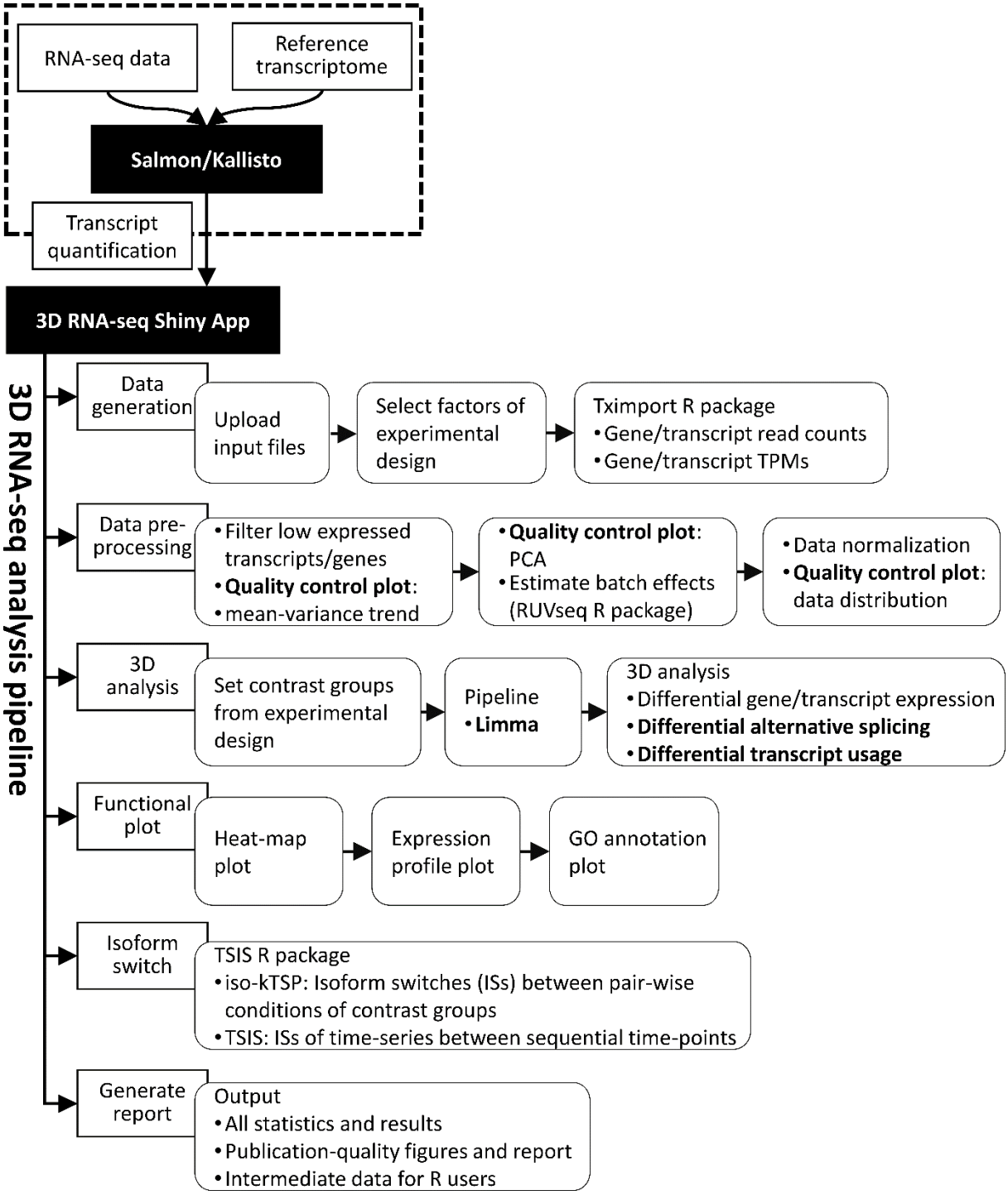


Figure 2. 3D RNA-seq analysis pipeline.

bioRxiv preprint doi: https://doi.org/10.1101/656686; this version posted May 31, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

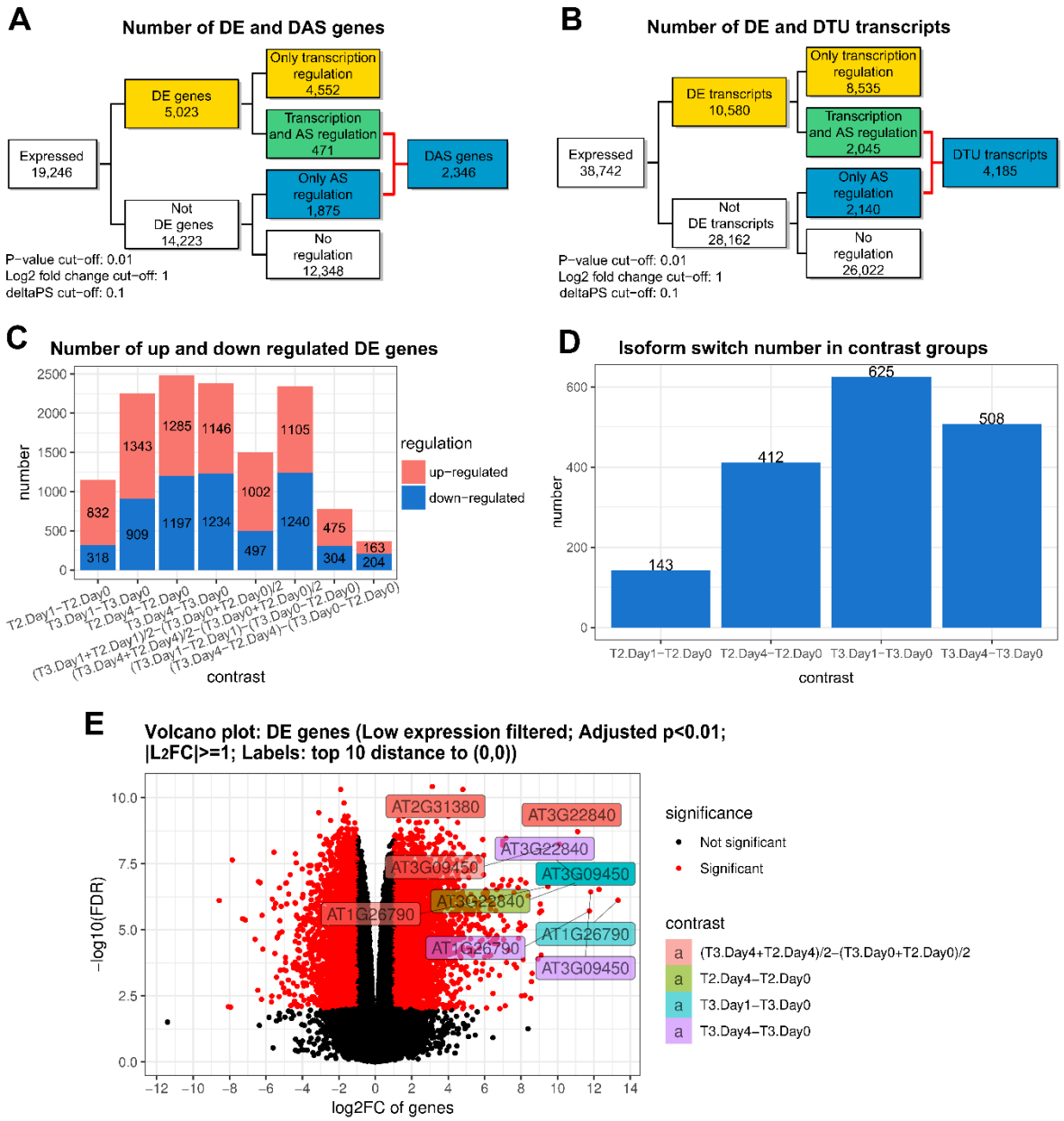


Figure 3. Illustrations of visualization outputs from 3D RNA-seq. A) Summary figure of expressed genes and significant DE, DE+DAS and DAS genes from analysis of the six time-points of the Arabidopsis data; B) Summary figure of expressed transcripts and DE, DE+DTU and DTU transcripts; C) Number of significantly up- and down-regulated DE genes in different contrast groups, and D) Number of significant isoform switches among contrast groups. E) Volcano plot of significant DE genes. The top 10 genes with the smallest *p* values and biggest fold changes are highlighted and different colours refer to different contrast groups.

bioRxiv preprint doi: https://doi.org/10.1101/656686; this version posted May 31, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

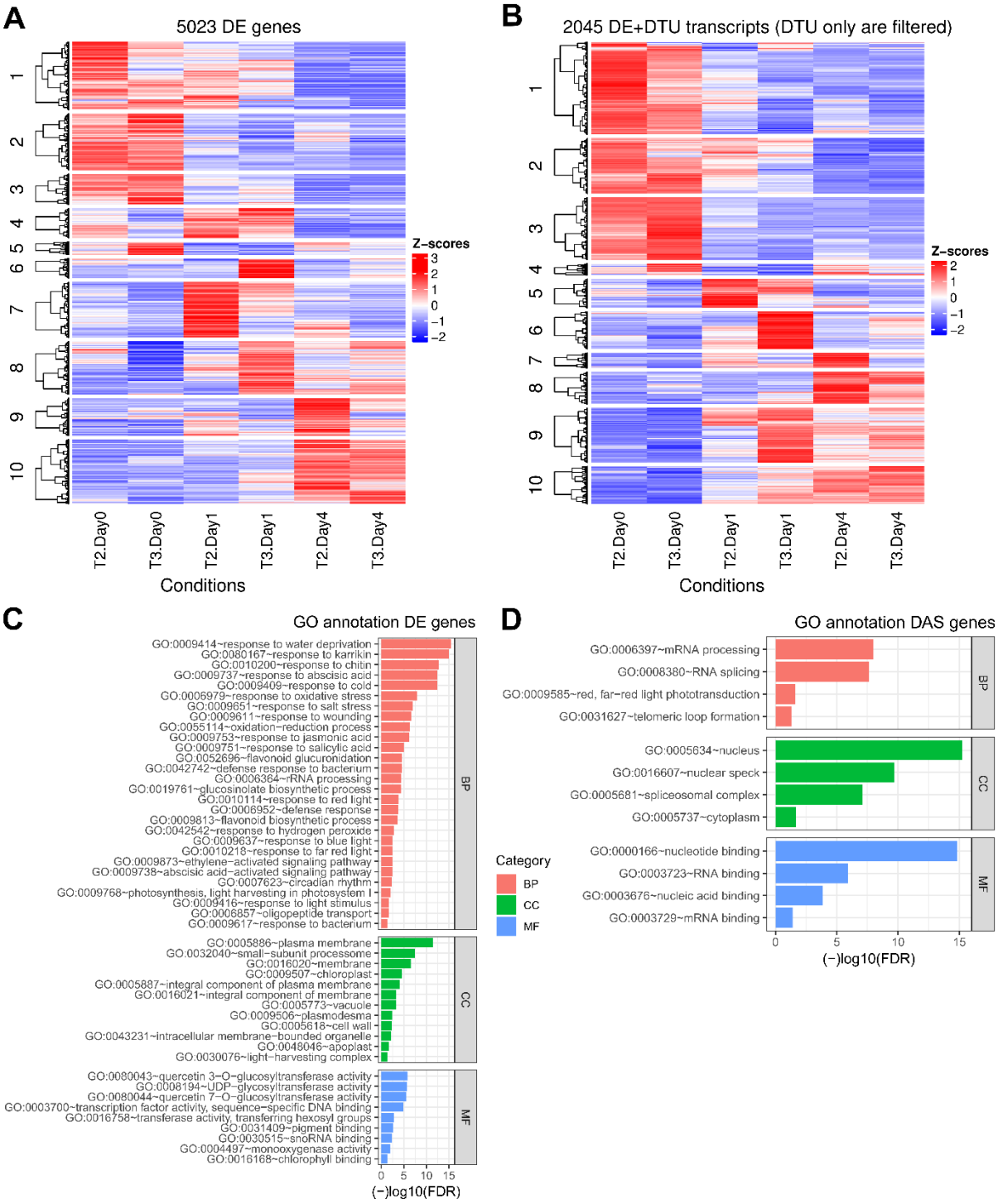


Figure 4: Downstream analyses of co-expression clusters and functional annotation. Heatmaps show the grouped expression profiles for A) DE genes and B) DTU transcripts across the samples. The top enriched GO terms for of C) DE and D) DAS genes are visualized with their associated FDRs.

bioRxiv preprint doi: https://doi.org/10.1101/656686; this version posted May 31, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

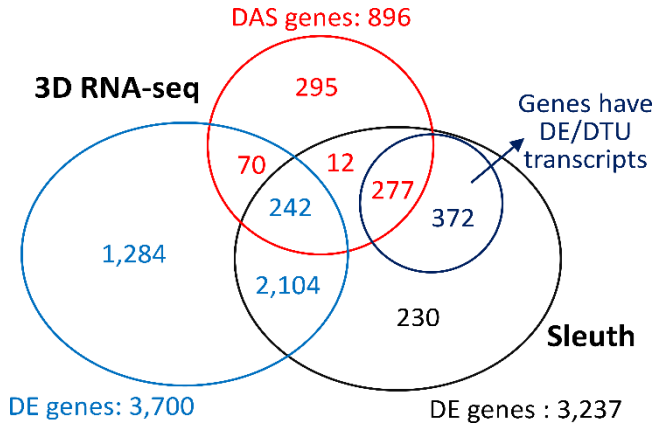


Figure 5. Comparison of the gene lists generated by 3D RNA-seq and Sleuth pipelines. The RNA-seq data on dexamethasone treatment of mice cells was taken from Frahm *et al*., (2018). Similar statistical parameters were applied when running 3D RNA-seq and Sleuth. The Venn diagram compares the DE genes from Sleuth to DE and DAS genes and DE and DTU transcripts from 3D RNA-seq.

bioRxiv preprint doi: https://doi.org/10.1101/656686; this version posted May 31, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

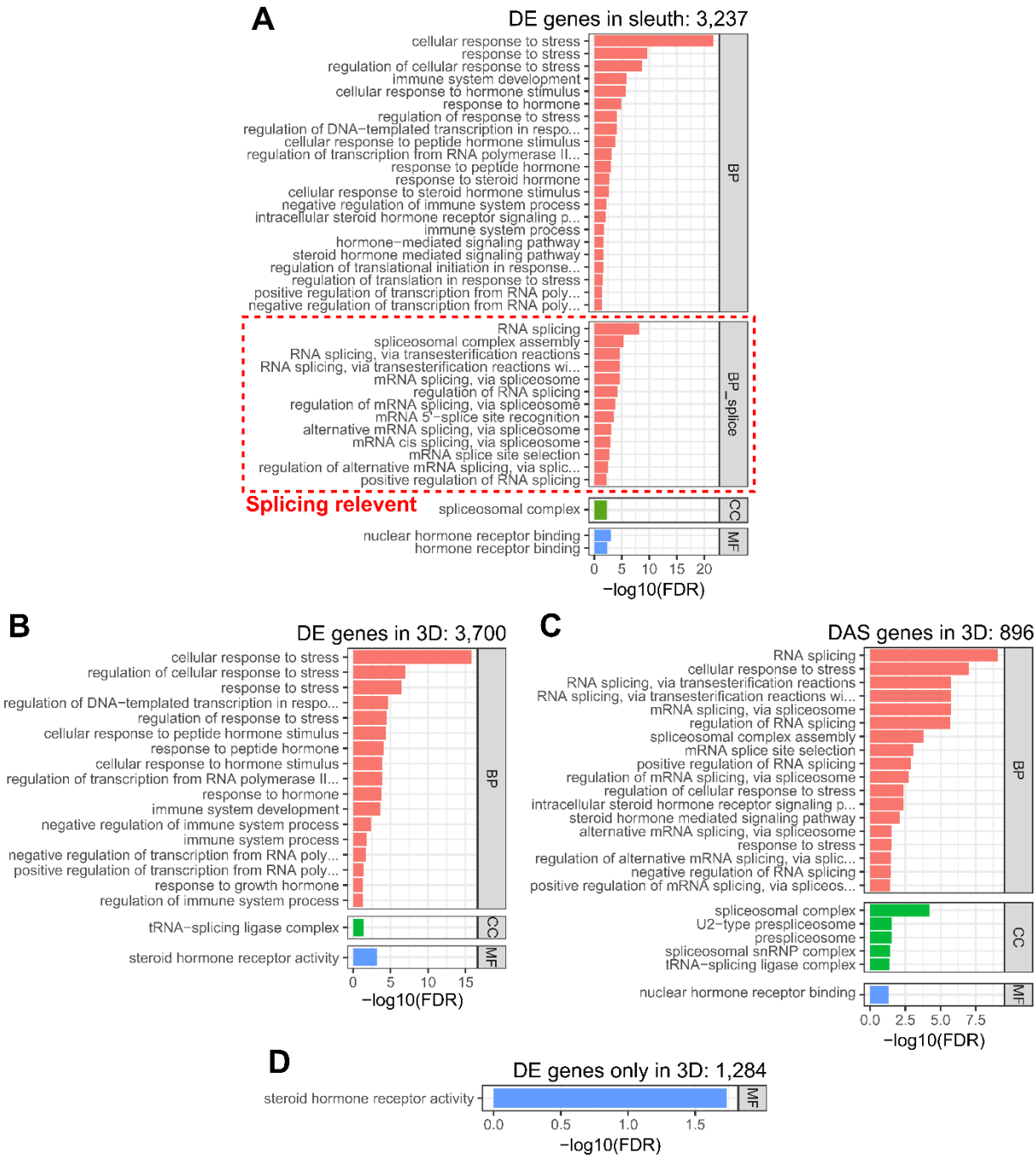
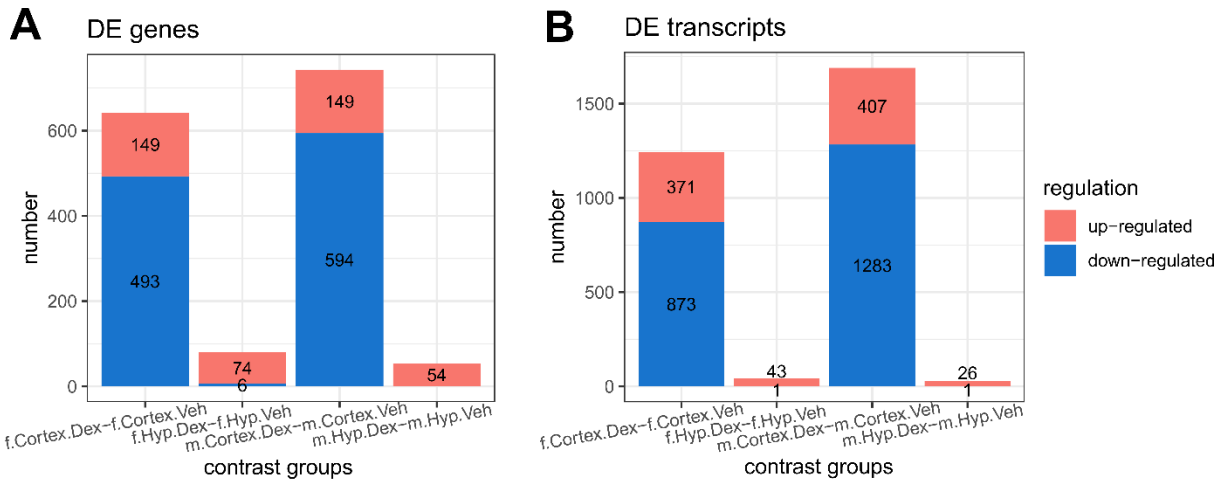


Figure 6. Top enriched GO terms on the significantly perturbed genes identified by 3D RNA-seq and Sleuth in the mouse RNA-seq data. The Fisher's exact test and topGO R package were used to generate significant enrichment gene ontology (GO) terms with FDR < 0.05. Only the GO terms of response to stress, immune system, inflammatory, hormone, splice, splicing, spliceosome and spliceosomal are presented. A) Significantly enriched GO terms of DE genes from Sleuth; B) Significantly enriched GO terms of DE genes from 3D RNA-seq; C) Significantly enriched GO terms of DAS genes from 3D RNA-seq; and D) Significantly enriched GO terms of novel DE genes unique to 3D RNA-seq. Splicing/spliceosome related

bioRxiv preprint doi: https://doi.org/10.1101/656686; this version posted May 31, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

GO terms are enriched in the DE genes in *Sleuth* (red dashed box in A) but are found in GO terms associated with DAS genes in 3D RNA-seq (C). . BP: Biological process; BP_splice: Biological process with terms of splice, splicing, spliceosome and spliceosomal; CC: Cellular Component; MF: Molecular Function.

bioRxiv preprint doi: https://doi.org/10.1101/656686; this version posted May 31, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.



C

Contrast \ Levels	Gene level			Transcript level		
	DE only	DE&DAS	DAS only	DE only	DE&DTU	DTU only
f.Cortex.Dex-f.Cortex.Veh	624	18	208	1179	65	198
f.Hyp.Dex-f.Hyp.Veh	79	1	5	43	1	20
m.Cortex.Dex-m.Cortex.Veh	716	27	336	1569	121	303
m.Hyp.Dex-m.Hyp.Veh	54	0	8	26	1	21

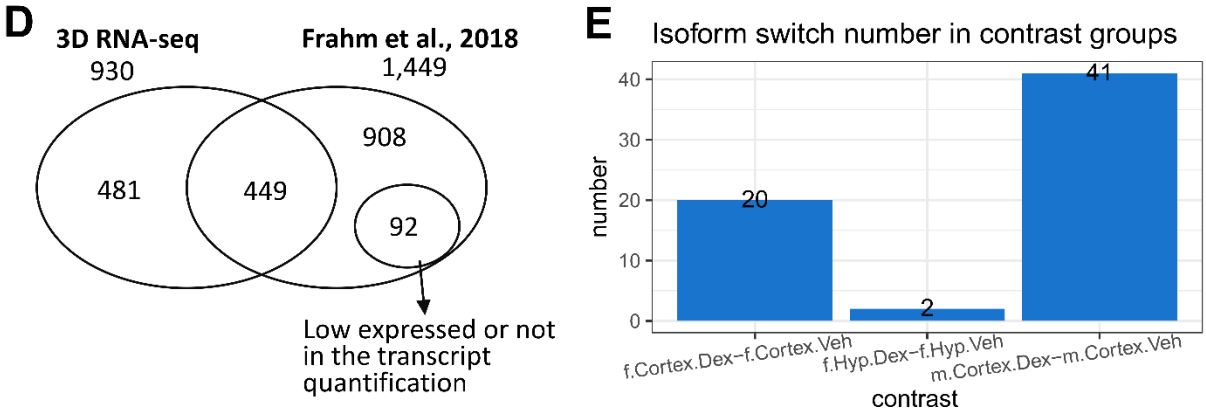


Figure 7. Sex-specific and tissue-specific 3D RNA-seq analysis of the mouse data. Contrast groups were designed to investigate Dex-induced expression and alternative splicing changes between male and female and cortex and hypothalamus brain regions. Significant DE gene/transcript lists were generated by BH adjusted p-value < 0.05, $L_2FC \geq 1$ and $\Delta PS \geq 0.1$. A) Up- and down-regulated DE genes and B) DE transcripts. C) Summary of statistical analysis results from 3D RNA-seq in each contrast group. D) Venn diagram comparing the DE genes in the 3D RNA-seq analysis to the results in Frahm *et al*., (2018) in which the significant DE genes were determined by p-value < 0.05 (multiple testing adjustment and L_2FC cut-off were not applied). 92 genes had low expression and were not included in the transcriptome quantification in 3D RNA-seq analysis. E) the number of Isoform switches in different contrast groups with the following cut-offs: probability of switch ≥ 0.5, difference of average TPMs at different conditions ≥ 1 TPM and adjusted p-value of the TPM difference < 0.05.