Development of a scoring function for finding differential expression - application to Lolium Perenne

Albert Fradera Sola
IBERS. Aberystwyth University
albertfradera85@hotmail.com

Narcis Fernandez Fuentes
IBERS. Aberystwyth University
naf4@aber.ac.uk

ABSTRACT

RNA-sequencing for detecting changes between expression patterns has emerged as a frequent tool in life sciences studies. However, it is an under-development tool which still lacks a standard procedure. This way, many software packages and approaches could be used when designing an experiment. Here we develop and present an approach based on a scoring function. It allows using multiple packages and relies its efficiency on validated data. Thus, the function is adjusted using qPCR verified data. Then it is tested on experimental data obtained from a *Lolium Perenne* drought stress tolerance study yielding positive results with several potential uses in further studies.

1. INTRODUCTION

As years go by, high-throughput cDNA sequencing (RNA-seq) has become a popular approach to transcriptome characterization. It allows transcript identification and differential expression assessment [43], being both key points in molecular biology. Moreover, RNA-seq has certain advantages over microarray techniques: it can be performed without prior knowledge of reference sequences [18] and allows transcriptome de novo assembly [36], quantification [22], and alternative splicing detection [3]. This way, RNA-seq has been proven a powerful and successful approach [42], which, added to its continuously cost decrease [26], makes it a frequent tool in life sciences research [32]. However, it is still an under-development tool so, currently, the procedure has not been standardized. Moreover, its widespread usage has led to an arise of different pipelines which specially differ at differentially expression assessment [37, 39, 46]. Even so, efforts have been made to stablish a survey for best practices [15]; the RNA-seq core analysis always include transcriptome profiling (alignment to a reference genome or *de novo* assembly) and differential expression (statistical methods applied to test the significance of differences between groups). In figure 1 we depict the general RNA-seq workflow.

Publication: To appear in the thesis defence of the MsC in Omics Data Analysis, September 2017, Barcelona.

Acknowledgments: The authors would like to thank the Aberystwyth University, the Institute of Biological, Environmental and Rural Sciences (IBERS) and the members of the bioinformatics core facility.

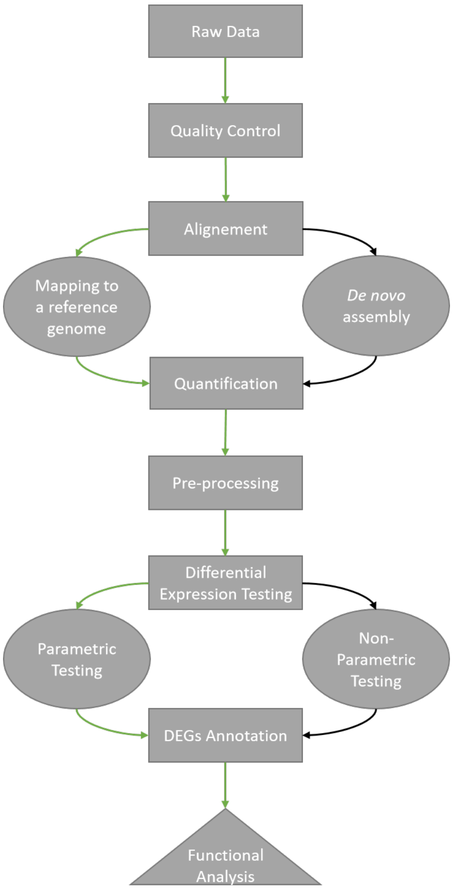

Figure 1: RNA-seq workflow schematic representation. Squares represent mandatory steps while triangles represent the optional ones. Circles represent different ways to perform steps. Green arrows point the procedure followed in this study.

Consequently, RNA-seq is used in a variety of different analysis scenarios and has multiple applications. One of them, and the case of study in this article, is the identification of differentially expressed genes (DEGs). It has some difficulties. Some are inherent to next-generation sequencing procedures; bias is introduced by variation in the nucleotide composition between genomic regions or by the larger read coverage that receive longer transcripts [25]. This 'inner-sample' kind of bias problem is usually ignored as it is considered to affect all samples in a similar way [29]. Thus, in early RNA-seq studies, samples fitted well to a Poisson distribution which assumes that the variance is equal to the mean [24]. But, due to its cost reduction [26], biological replicates have become available. So, a new phenomenon, called overdispersion, is introduced making the Poisson distribution to underestimates data variation which results in a higher false positive rate [34]. Thus, methods based on negative binomial model were introduced as they deal better with biological variability and overdispersion [2, 19, 35]. Overall, we have a small constellation of similar statistical methods oriented to the same purpose: to determine the DEGs. However, even if they approach the same question, every method uses their inner way to get an answer; and the problem is that it is not usually convergent [37, 39, 46]. We have, therefore, a major dilemma: the way in which we face the problem determines the result that we will obtain [37, 39, 46].

The aim of this study is to reduce the impact of this problem, in order to maximize and strengthen the detection of DEGs. Consequently, we will evaluate the performance of four widely used statistical software packages (baySeq [19], DESEq2 [2], edgeR[35], lima - voom transformation[20, 38]) and build a scoring function around them. This scoring function is tuned through a validated data set ('gold standard') with the aim of determining which genes are truly DEGs. Then is tested with an experimental data set, which, in addition, is used for a complete RNA-seq analysis. It seeks to observe how affects drought conditions to the growing of the grass *Lolium Perenne*. Therefore, it is intended to provide one more tool for the design of an RNA-seq experiment and evaluate its results by performing a *Lolium P.* drougth stress tolerance study.

2. MATERIALS AND METHODOLOGY

Unless specified, ongoing procedures are only applied to the experimental data set and performed with R software [30]. Since we directly download cured count matrices, preliminary procedures do not need to be performed on the validated data set.

2.1 Data sets

2.1.1 Validated data set

For the gold standard, we use data presented in the Rajkumar et al. 2015 study [31]. Its experimental design is depicted in figure 2. In their study, they obtained RNA samples extracted from amygdalae micro-punches of a genetically modified mouse strain (Brd1+/-) and of their wild-type (WT) littermates (8 biological replicates per group) [31]. After sequencing, they used 4 methods (Cuffdiff2 [44],DESeq2, edgeR and TSPM [4]) to obtain each DEGs list [31]. Then, they randomly selected 115 genes from those who were identified as DEGs by the four methods. Finally, they validated them using independent biological replicates and high-throughput quantitative reverse-transcription PCR (qPCR) [31]. Among data they present, which is contrasted and public available, we can find the counting matrices they used for the DEGs analysis. Thus, we obtain and use them for our own DEGs analysis. We also download a file containing the randomly selected 115 genes. It indicates whether they are differentially expressed or not, according to the qPCR validation.

2.1.2 Experimental data set:

For the experimental data set, we use our own data; it comes from growing the grass *Lolium Perenne* under four different conditions, as depicted in figure 2. These are based on the humidity of the soil, using a factor known as soil water content (SWC) and are intended to simulate drought conditions. This way, the lower is the SWC the drier the soil, allowing us to observe its effects on *Lolium P. growth.* Thus, plants were grown in compost initially, single tillers taken, roots cut close to base of plant, rinsed of compost and transferred to containers of water until they showed new root growth on a controlled environment at 20°C and 8 hours photoperiod. After about 6 days on average they were put into 90 mm pots of vermiculite (graded for horticultural use 2.0-5 mm) to establish and watered with hoaglands twice a week. Once established - between 15 and 21 days from tillering, watering was stopped and SWC was monitored using a moisture meter HH2 Delta-T meter (A Delta-T devices). Leaf and root samples from same clones were samples and 35%, 15%, 5% and 1% SWC moisture levels were reached. Then, we collect samples from two different tissues (root and leaf) of each replicate. Leaf samples were cut and flash frozen in liquid nitrogen and stored at -80°C. The roots were washed with distilled water and blotted dry prior to storage also at -80°C. At this point, we have 16 samples from each tissue (4 for each SWC) ready for sequencing.

2.2 Sequencing:

Total RNA was extracted using Trizol reagent (from Thermo Fisher) following the instructions provided by the vendor. RNA extracts were then cleaned using a Qiagen RNeasy MinElute column (cat no 74204). Then, samples were quantified and send to IBERS genomic facilities for sequencing. The pair end libraries were prepared in accordance to the standard protocol provided by Illumina for the HighSeq sequencer, which returns us the raw data in .FASTQ [13] format. It is a compressed file which contains read data from the sequencing process. Thus, it contains both the sequence and an associated per base quality factor [13].

2.3 Quality control and Alignment:

Starting from the raw data, we perform a quality control following these general criteria: (1) filter truncate reads (remove nucleotides at the beginning/end of each read). (2) filter trim adapters (remove nucleotides at the beginning and/or at the end of each read that match the adapter sequence). (3) filter low quality reads (GC content, PCR artifact, overrepresented k-mers. . .) [15]. Once the quality control is done, we align our reads to a *Lolium P.* reference genome [12]. Results are cointained in .BAM files [23].

2.4 Quantification:

The code used for obtaining the count matrices can be found at supplementary file 1. We want to evaluate differential expression among genes under our specific conditions so we need to know the number of reads that map to each transcript sequence. This can be done at several levels (transcript, exon, intron.) but, since our goal are genes, we do

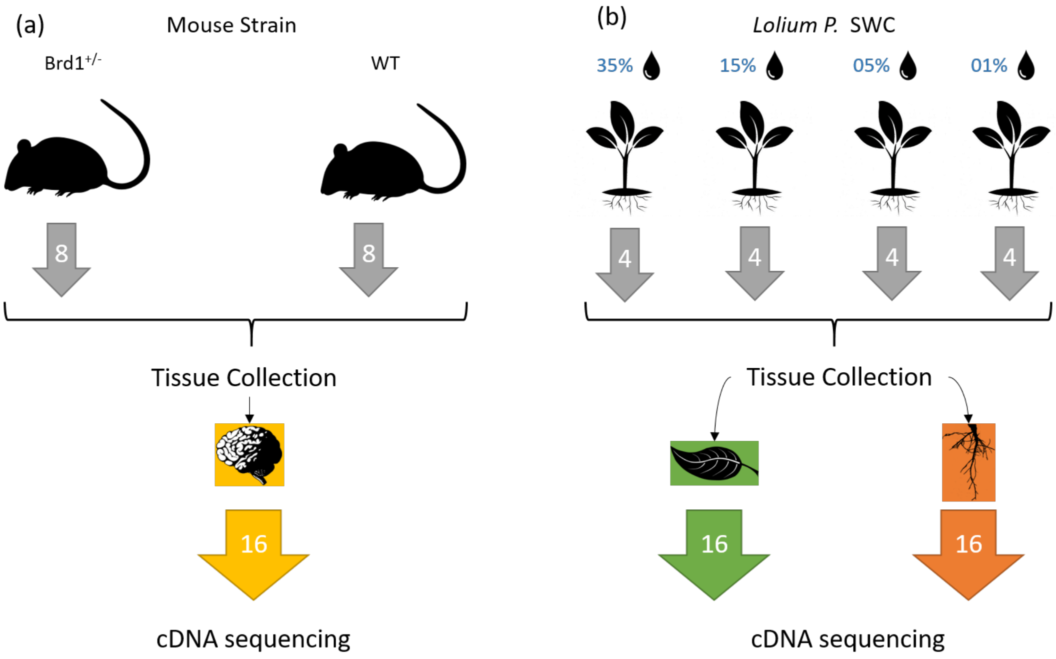

Figure 2: Data sets experimental design. (a). Validated data set design. In this case, we have two groups consisting of 8 individuals for each experimental condition ($Brd1^{+/-}$ knockouts and wild type). From each individual, a micro-punch of amygdala is extracted making a pool of 16 samples ready for sequencing.(b) Experimental data set design. In this case, we have four groups consisting of 4 individuals for each experimental condition (a 35%, 15%, 05% and 01% of Soil Water Content). From each individual, a sample of leaf and root tissue is extracted, making a pool of 16 samples for each tissue ready for sequencing.

it at gene level. Thus, we perform the quantification by relating our BAM files (which contain information about each read and its genome coordinates [23]) to a reference genome [12]. Our *Lolium P.* reference genome [12] is contained in a .GTF file, which contains the genome coordinates of exons and genes [15]. Then, using the R packages GenomicFeatures and GenomicAlignments [21] we build our counting matrices. They contain the number of counts per annotated gene for each sample.

2.5 Pre-Processing

The code used for the preprocessing for both root and leaf samples can be found at supplementary file 2.

2.5.1 *Independent filtering:*

Our first step, is to filter our counting matrices and eliminate all those genes who have less than 1 count within all samples [10, 40]. The main reason for this filtering concerns to our statistical power. Later, we will be performing many statistical tests with its consequently test correction. By omitting those genes that have little or no chance of being detected as differentially expressed, we avoid some statistical power loss [10]. This way, we filter out unexpressed genes keeping only genes that are expressed in at least one sample.

2.5.2 *Multivariate visualization and ordination:*

It is also crucial to look for batch effects and to assess the global quality of our RNA-seq data. Hence, we are seeking for samples whose experimental treatment suffered from abnormalities that could make data points obtained from them to undermine our global quality [15]. Before starting with the procedures, and because count data is heteroskedastic, we apply the regularized logarithm transformation (rlog) that comes with the DESeq2 software package [2]. If the data is used on the original count scale, the result will be dominated by highly expressed, highly variable genes. Thus, we use the shrinkage approach of DESeq2 to implement rlog transformation making data more homoscedastic. It behaves similarly to a log2 transformation for genes with high counts, while shrinking together the values for genes with low counts[2]. This way, we manage to avoid the spreading of data by creating a similar dynamic range [2]. Therefore, multivariate visualization and ordinations tend to work better [2]. Once data is transformed, we perform the following plots and analysis which help us to look for between-samples biases: (1) Box and density plots. By observing the distribution of rlog counts we can contrast the distribution of gene-level expression values on different samples. (2) Principal component analysis (PCA). PCA is used to reduce multidimensional datasets to lower dimensions for analysis; it is a technique that can determine the key features of high-dimensional datasets. In other words, it gives a view of the correlation of expression between samples: data is projected on several axes (or components), ordered by decreasing significance [6]. This way we can express the maximum variation with the minimal variables. When plotted, it is useful for visualizing the overall effect of experimental covariates and batch effects. In the context of RNA-Seq

Table 1: Features regarding the chosen software packges.

Software Package	Normalization Procedure	Distribution Assumption	Statistical Test
baySeq	Quantile scaling factors	Negative Binomial	Posterior Likelihood Comparison
DESeq2	RLE	Negative Binomial	Wald Test
edgeR	TMM	Negative Binomial	quasi-likelihood F-test
limma - voom transformation	TMM	voom data transformation	Empirical bayes testing

analysis, PCA essentially clusters samples by groups of the most significantly dysregulated genes. Clustering first by the most significant group, then by progressively less significant groups. Given the experimental design of the dataset that we are attempting to analyse here (e.g., samples belong to four distinct groups), there should be a clear separation of the groups of samples by the first components. So, biological replicates of the same condition will cluster together [6]. (3) Heat map of sample distances. In a similar way, and to explore the similarities and dissimilarities between samples, it is often instructive to look a heatmap of sample-to-sample distance matrix. There are several methodologies to compute distances; in this study, we stick to Euclidian distances

Once the multivariate visualization and ordination is performed, we can evaluate if there are samples that cause batch effects being this way considered as outliers. Normally it is clear (e.g. a sample which does not cluster with any condition in PCA); although, sometimes, there are samples which are not clear whether to be considered as outliers or not. In these cases, it is useful to calculate the cook's distance among least-squares regression analysis of the two first PCA variables. Those who are above four times the mean can be considered as outliers [16].

2.5.3 *Normalization procedures:*

The code for normalization procedures can be found at supplementary files 3, 4, 5 and 6. Our final step for reducing bias introduced by batch effects is normalization. The overall strategy is to choose an appropriate baseline, and express sample counts relative to that baseline. This way, normalization procedures take into account which is the sequencing depth and the inner heterogeneity of count distributions. If not, highly and differentially expressed features could skew the distribution [11]. Each software package chosen for this study has its own approach to achieve the normalization goal: (1) Quantile scaling factors [11], used in baySeq. (2) RLE [2], used in DESeq2. (3)TMM [33], used in edgeR and limma - voom transformation. They are also resumed in table 1. Despite these sample-specific normalization methods, batch effects may still be present.

2.6 Differential Expression Testing

The code for defferential expression testing procedures can be found at supplementary files 3, 4, 5, 6 and 7. After the pre-processing, data is ready for the statistical analysis. We perform the differential expression testing for both the validated and the experimental data set. This way, for de validated data set we will be testing if there is an expression pattern change between $brd1^{+/-}$ strain and their wild type littermates (figure 2). On the other side, for the validated data set we will be testing if there is an expression pattern change among 4 SWC levels which leaves room for several comparisons. Hence, we decide to apply two approaches: (1) testing against reference. We fix the 35% of SWC as our base level (it is the further level from drought conditions) and compare it against the other three levels. (2) testing against time course. In this case, each SWC level is compared to the one directly below it. Both approaches are depicted in figure 3. So, the validated data set and both approaches used for the experimental data set are tested with four software packages which we proceed to present. Its principal features are depicted in table 1.

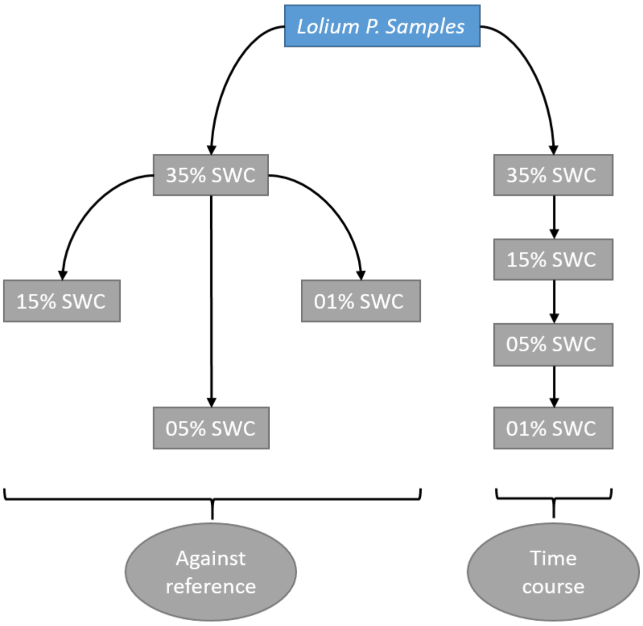

Figure 3: Differential expression testing approaches for Lolium P. leaf and root samples.

2.6.1 *baySeq:*

baySeq is presented in the 2010 Hardcastle *et al.* study [19]. It is based on estimating posterior likelihoods of differential expression via empirical Bayesian methods [19]. It assumes a negative binomial distribution [19]. Instead of regular significance values, it returns posterior probabilities and reports a Bayesian false discovery rate (FDR) estimate [19]. Thus, probabilities for differential and non-differential expression are computed and compared [19].

2.6.2 *DESeq2:*

DESeq2 is presented in the 2010 Anders and Huber study [2]. It assumes a negative binomial distribution [2]. Its modeling is based in the observed relationship between mean and variance [2]. Once data is modelled, a Wald test is performed for significance testing: the shrunken estimate of log fold change (LFC) is divided by its standard error, resulting in a z-statistic, which is compared to a standard normal distribution [2]. It returns a significance value whom FDR is controlled using the Benjamini-Hochberg procedure [8].

2.6.3 *edgeR:*

edgeR is presented in the 2009 Robinson *et al.* study [35]. It also assumes a negative binomial distribution [35]. In this case, differential expression is assessed using an empirical Bayes estimation and quasi-likelihood F-tests [35]. It returns a significance value whom FDR is controlled using the Benjamini-Hochberg procedure [8].

2.6.4 *limma - voom transformation:*

limma - voom transformation is presented in the 2014 Law *et al.* study [20]. It also borrows tools from limma [38] (its microarray-based sibling) and edgeR [35] whose normalization procedures are required. Unlike the other packages, it is not based on negative binomial distribution. It transforms count data to logarithmic (base 2) scale and estimates their mean-variance relationship seeking for linear modeling. Finally, Empirical Bayes testing is performed [20]. It returns a significance value whom FDR is controlled using the Benjamini-Hochberg procedure [8].

2.6.5 *Scoring function:*

From each software package, we obtain a list of genes with an associated significance value. We could use that value to set a threshold and decide whether a gene is differentially expressed or not. However, we choose a different approach and decide which genes are differentially expressed by building a scoring function around the significance values as depicted in figure 4. The aim of the scoring function is to use the validated data set to find a threshold from whom we can decide which genes from the experimental data set can be considered as DEGs. The first step is to convert significance values to a unit that allows us to compare them. We choose to transform all significance values to z-scores [7]. It is the difference between a significance value and the mean for that significance value divided by the standard deviation for significance value as seen in equation 1.

$$Z = \frac{Y - \mu}{\sigma} \tag{1}$$

Where Y represents a significance value, μ represents its mean and σ its standard deviation.

With this transformation, we do not only make the different values of significance obtained comparable to each other but make them independent of their population size [7]. Once we have our z-scores, we are ready to find and set our significance threshold. We try four definitions for our scoring function: (1) The sum of z-scores, as depicted in equation 2. (2) The mean of z-scores, as depicted in equation 3. (3) The median of z-scores, as depicted in equation 4. (4) The variance of z-scores, as depicted in equation 5.

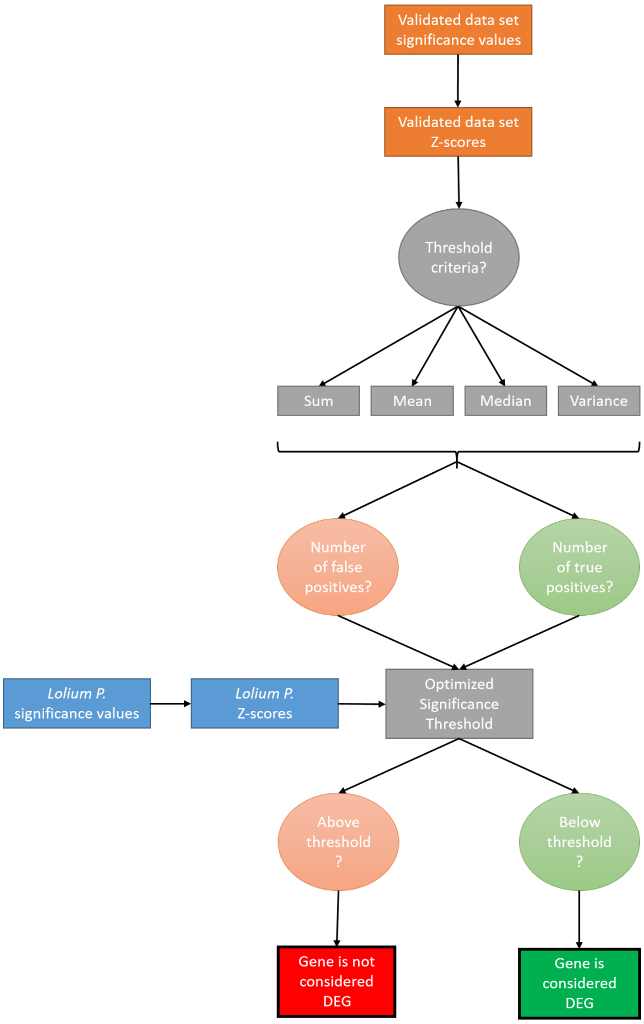
Validated data set significance values
Validated data set Z-scores
Threshold criteria?
Sum
Mean
Median
Variance
Number of false positives?
Number of true positives?
Lolium P. significance values
Lolium P. Z-scores
Optimized Significance Threshold
Above threshold ?
Below threshold ?
Gene is not considered DEG
Gene is considered DEG

Figure 4: Diagram depicting how the scoring function works. In brown, we can find the validated data and its role in fixing the optimized threshold. In blue, we can find the experimental data and how the previous threshold is used for determining the DEGs.

After applying the equations to the validated data set, we obtain a list with genes and their associated four scores. Using their qPCR validation study [31], we know from 115 genes who are whether true positives (detected as DEG and validated with qPCR) or false positives (detected as DEG and not validated with qPCR). We mark them and rearrange the list until we find which function and score allows us to have the higher precision and recall [5].

$$Z_{sum} = \sum Z_{scores} \tag{2}$$

Where Z represents each gene z-score obtained from each software package.

$$Z_{mean} = \frac{\sum Z_{scores}}{n} \qquad (3)$$

Where Z represents the z-score of each gene and n is the total number of z-scores.

$$Z_{median} = O_{(n+1)/2}, \qquad (4a)$$
$$Z_{median} = (O_{N/2} + O_{N/(2+1)})/2, \qquad (4b)$$

Where O represents the sorted z-score of each gene and n is the total number of z-scores. In case that n is odd, equation 4a is used. Otherwise, we use equation 4b.

$$Z_{variance} = \frac{\sum (Z_{scores} - Z_{mean})^2}{n-1} \qquad (5)$$

Where Z represents the z-scores of each gene, z-mean represents its mean and n is the total number of z-scores.

At this point, we know which scoring function and score performs better; we set it as our optimized significant threshold. Then, we transform the significant values from our experimental data set to z-scores. We know which function to apply, thanks to our previous work with the validated set. We apply it, and, using the optimized significance threshold, assess which genes are to be considered DEGs. The whole process is depicted in figure 4.

2.7 DEGs annotation and functional analysis

Finally, we want to annotate our results and perform a functional analysis. We use the Blast2GO tool [14] for both purposes. Our first step is to retrieve the sequence and genome coordinates from our DEGs. Then, the program annotates the sequences using the BLAST[1] and interPro [17] algorithms. Once we have the annotated DEGs, the program continues with the functional analysis for which relies on Gene Ontology (GO) database [9]. Thus, each gene gets its biological context from the three ontologies (biological process, molecular function, cellular component) included in the database making its functional annotation complete. Our last step is to run a fisher's exact test [45] to find which GO terms are overrepresented in our data set compared to the reference genome [12]. Then, they are depicted using Revigo plots [41]; its goal is to summarize long, unintelligible lists of GO terms by finding a representative subset of GO terms using a clustering algorithm that relies on semantic similarity measures[41]. Among other values, it returns a uniqueness and a dispensability value. The first one measures how unique is a term when semantically compared to the whole list. On the other hand, dispensability sets the threshold at which a term is removed from the list or assigned to a cluster [41]. Thus, a reduced list of GO terms can be depicted and visualized in function of their semantic relatives. Revigo plots are based on multi-dimensional scaling, as seen before with PCA. This way, GO terms dimensionality is reduced and those which are semantically similar cluster together. In this case, axes have no intrinsic meaning. The plots also allow to show the GO term significance, in terms of overrepresentation.

Once the whole process concludes, we end up identifying which genes are differentially expressed and with a knowledge of their biological context. Hence, we are in strong position for evaluating the effects of drought stress in *Lolium P.* expression pattern.

3. RESULTS AND DISCUSSION

3.1 Pre-Processing

We start the pre-processing analysis with the count matrices obtained from the quantification process. Once, the independent filtering is done we proceed to check for abnormalities on the sample-sample relation, both for leaf and root samples.

3.1.1 Experimental data set - Leaf samples:

All 16 samples from the leaf tissue behave as expected. Observing the box and density plots we can state that there is no sample with an unusual counting behaviour. Both plots can be found at supplementary file 2. Also, PCA analysis and heat map of sample distances show great results. In the PCA plot, depicted in figure 5 (a), we can see how samples cluster by condition. Thus, each sample fall within its SWC ellipsis and there is none who appears separately. The motivation behind the PCA is to find to find new variables that explain as much of the variability as possible. Each new component results from the sum of the initial variables [6]. Thus, PCA components correspond to the direction of greatest variability [6]. Moreover, within the first two components of the PCA the 69% of data variance is contained.This way we can state that: (1) leaf samples variability is driven by SWC and (2) most of the data set variability is explained by the SWC.

Heat map results, depicted in 5 (b), show results in the same direction. The closer is the Euclidian distance between two samples, the bluer is their relation. This way, samples from the same condition show the deeper blue and fall in the same branch of the hierarchical cluster dendrogram.

Finally, we observe that 01% samples are further from the others both in the PCA and the heat map. Variability reflects expression patterns, so we have the first indicator that in drought conditions *Lolium P.* leaf tissue change its expression pattern.

3.1.2 Experimental data set - Root samples:

Contrary to what we see with leaf samples, root data analysis does not have a straightforward lecture. Counting plots, which can be found at supplementary file 2, show the expected results. But it is in the PCA analysis and heat map when samples start to behave oddly. If we observe the PCA components, depicted in figure 5 (c), results go in the same direction that leaf samples. Most of the variability is explained by the SWC condition (in this case it goes up to the 86%). However, we can see that not all samples fall with its condition. Even we can suspect of samples being outliers because they fall far apart from they own kind.

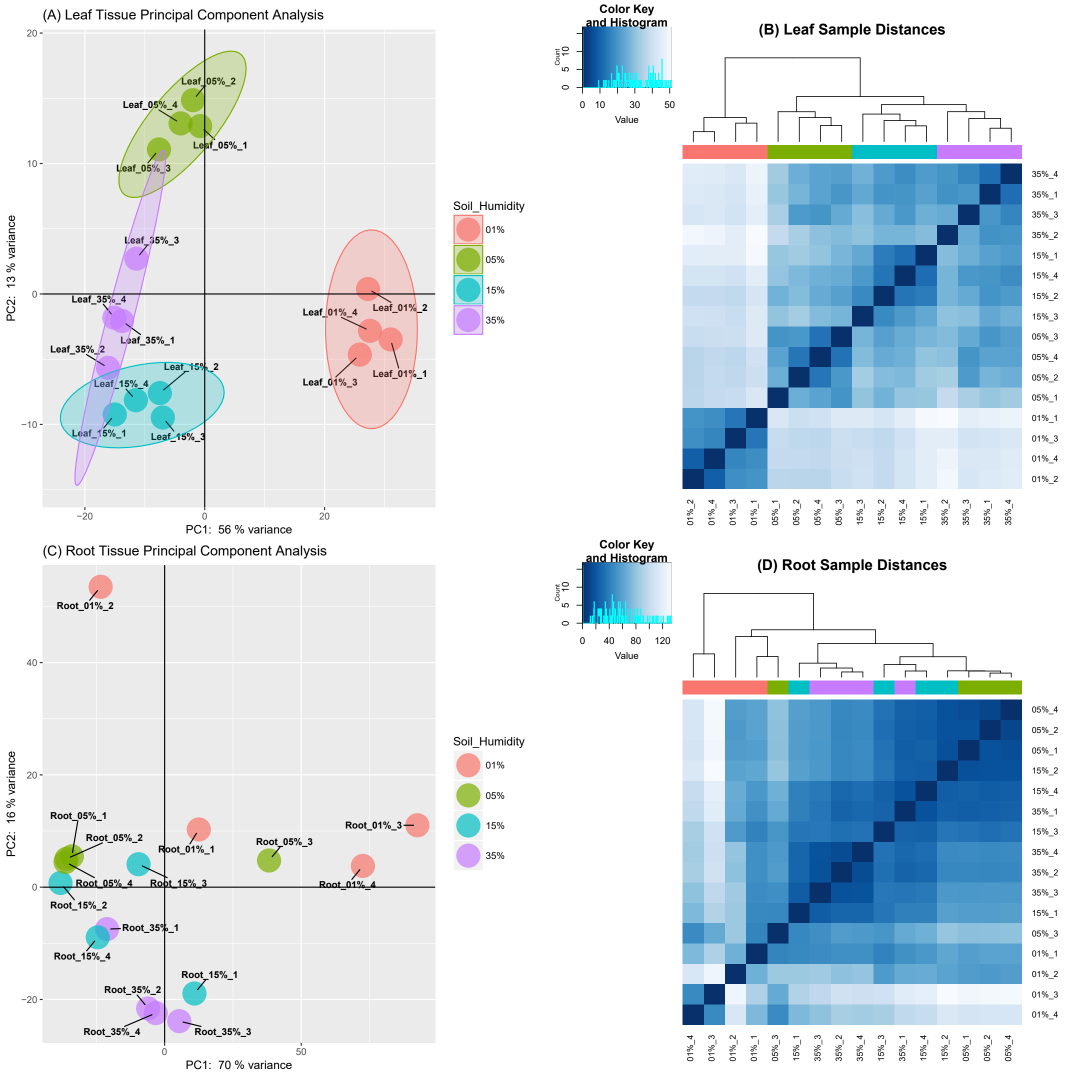

Figure 5: Pre-processing results of leaf and root samples.(a) Leaf samples PCA plot. (b) Heat map of leaf sample distances.(c) Root samples PCA plot. (d) Heat map of leaf sample distances. Color code: purple for 35%, blue for 15%, green for 05% and red for 01% level of Soil Water Content.

Heat map results, depicted in figure 5 (d), show results in the same direction. Samples do not cluster well in the dendrogram. We can also see that there are more clusters than conditions, and samples which fall within wrong condition clusters. Even so, by only observing the plots we cannot state that bad clustering answers to an outlier situation. This way, we decide to compute the cook's distance among least-squares regression analysis of the two first PCA variables. Results are depicted in a plot, which can be found in supplementary file 2, and in table 2. We can observe that replica 1 for 35% is close to reach the four times the mean threshold, but it does not. So, none of the samples have a bigger than four times the mean cook's distance. Thus, we decide that none of the samples should be considered an outlier and that oddities are not due to batch but experimental effects. Consequently, none of the samples is removed from the analysis and it proceeds even with the multivariate and ordination oddities.

Table 2: Cook's distance among least-squares regression analysis of the two first PCA variables for each experimental condition.

	35%	15%	05%	01%
Replica 1	84.50	1.98	0.15	0.35
Replica 2	0.19	0.26	0.04	0.01
Replica 3	0.62	1.11	0.43	0.26
Replica 4	0.06	0.12	1.26	0.01
Mean (4x)	85.37	3.47	1.87	0.63

3.2 Differential Expression Testing:

From each analysis, we obtain a list of genes with an associated corrected significance value. Those lists are at supplementary files 8, 9, 10, 11 and 12. This way, we could infer from this value whether a gene is considered as differ-

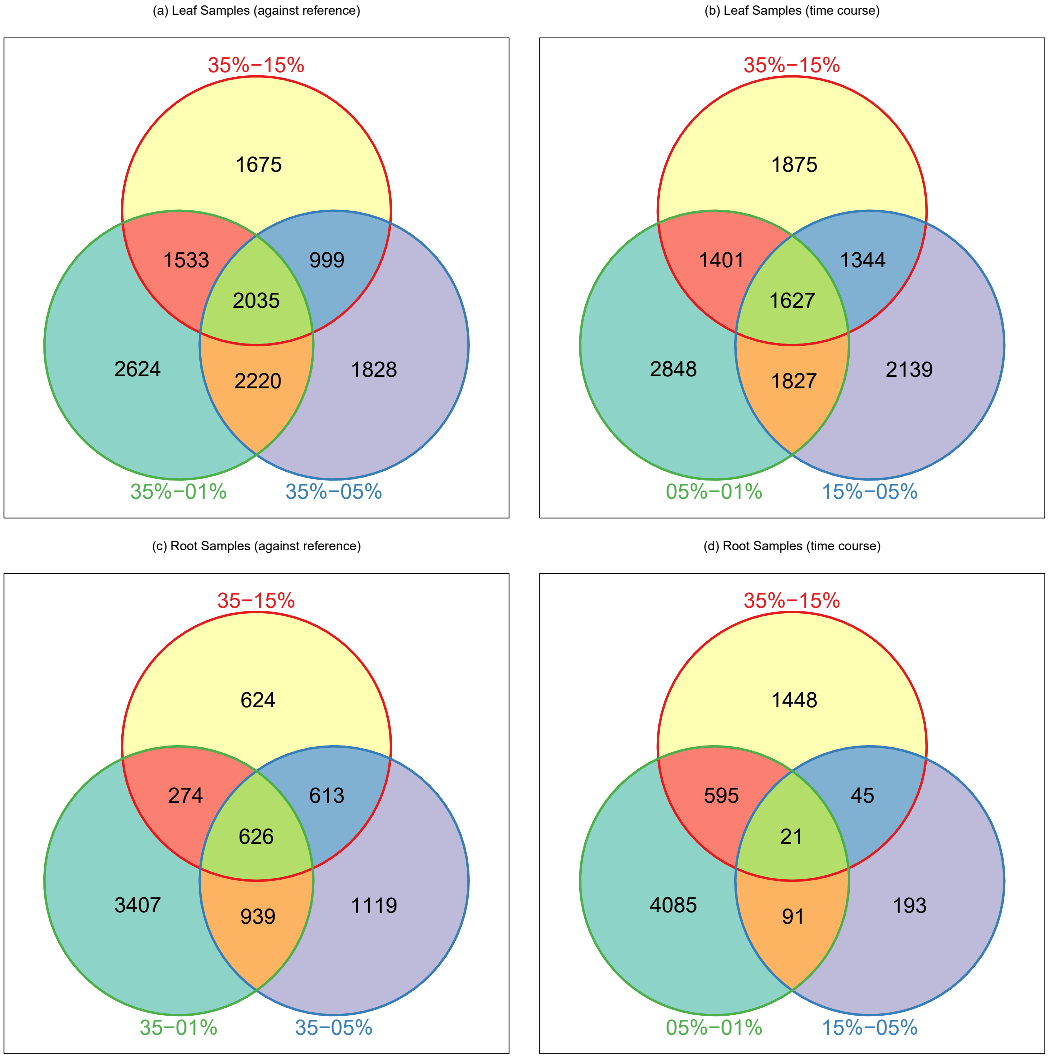

Figure 6: Venn Diagrams of DEGs.(a) Leaf samples (against reference). (b) Leaf samples (time course).(c) Root samples (against reference). (d) Root samples (time course).

ently expressed or not by setting a significance threshold. However, we use a different approach which is assessing the significance with the scoring function. Thus, we apply the z-transformation to each gene's significant values. On average, each gen has four associated significant values, one per each software package. Although there are few genes that have less than four related significant values. That is caused by the internal gene outlier detectors from each procedure, which cannot be skipped. Those detectors are not coincident making some genes to have less than four related significant values. Consequently, they also have less z-scores. Once we have the z-scores we proceed to apply the scoring functions, to both the validated and experimental data set.

3.2.1 *Validated data set:*

Data resulting from applying the z-score transformation and the scoring function to the validated data set can be found at supplementary file 8. The goal on using a validated data set is to set an optimized significant threshold (figure 4). After highlighting the true and false positives and rearranging the list using the four scoring functions, we notice that the mean one (equation 3) performs better than the others. We have a population of 115 genes, 60 from which are true positives and 55 which are false positives. Using the mean of z-scores as scoring function and fixing a threshold of -0.75, we detect 50 out of 60 true positive genes (83.3%) and 8 out of 55 false positive genes (14.54%). Thus, we have a great precision and recall which leads us to set a z-score mean of -0.75 as our optimized significance threshold. This way, the purpose of the validated data set is accomplished, and no further details (e.g. which genes are detected, biological implications.) are commented.

3.2.2 *Experimental data set - Leaf samples:*

Data resulting from applying the z-score transformation and the scoring function to the leaf samples can be found at supplementary files 9 and 10. It is also resumed in table 3, which depicts the total number of DEGs detected.

Table 3: Number of Leaf DEGs detected on each statistical approach (against reference or time course)

Leaf					
Against Reference			Time Course		
35%-15%	35%-05%	35%-01%	35%-15%	15%-05%	05%-01%
6242	7082	8412	6247	6937	7703

We observe that using the time course approach we detect less DEGs than the ones detected with the against reference approach. Talking from the point of view of physiological context, comparisons in the time course approach are closer. This way, it is expected that the expression pattern suffers less changes, which is what we are observing. Following the same argument, it makes sense that the 35%-01% comparison (highest change in SWC) is the one with the highest number of DEGs.

We also draw Venn diagrams (figure 6 a and b) to evaluate how DEGs are distributed among comparisons. Thus, we can observe how many genes are shared among different comparisons and how many are not. There is no comparison whose DEGs are completely included in the others. Moreover, each comparison has its own pool of DEGs which are independent from other contrasts.

Overall, we observe how, in both approaches, the total number of DEGs increases as the SWC drops. This states that there is differential expression pattern in *Lolium P.* leaf tissue. We have to make sure that this change of expression pattern is related to the drought stress conditions which is revealed in the functional analysis part.

3.2.3 Experimental data set - Root samples:

Data resulting from applying the z-score transformation and the scoring function to the leaf samples can be found at supplementary files 11 and 12. It is also resumed in table 4, which depicts the total number of DEGs detected with each statistical approach.

Table 4: Number of Root DEGs detected on each statistical approach (against reference or time course). Abnormal results are coloured in light red.

Root					
Against Reference			Time Course		
35%-15%	35%-05%	35%-01%	35%-15%	15%-05%	05%-01%
2137	3297	5219	2109	350	4792

The pattern observed in leaf samples, where against reference has more DEGs than time course is not that clear in root samples. Thus, not all time course contrasts have less DEGs than against reference contrasts. There is one comparison which clearly breaks the pattern by showing an abnormally low number of DEGs. It is the time course comparison 15-05%, and we obtain 350 DEGs. Back to the preprocessing results (figure 5), we already saw that root samples do not cluster well; so, it is not surprising that expression pattern between close samples, such as 15% and 05% show less DEGs. We also draw Venn diagrams (figure 6, c and d) and drive similar conclusions than with leaf samples. Overall, despite punctual differences, we also observe that the number of DEGs increases as the SWC drops. So, there is differential expression pattern in *Lolium P.* root tissue.

3.3 DEGs annotation and functional analysis

Data resulting from the functional analysis can be found at supplementary files 13, 14 and 15. In this section, we only present and discuss data resulting from leaf samples, time course analysis. The other three approaches are currently under processing, so their results cannot be presented.

3.3.1 DEGs Annotation:

Using our DEGs sequences and the Blast2go tool, we seek for homology using the BLAST and interPro algorithms. Unfortunately, not all our sequences present homology; for all three comparisons, we manage to annotate around the 82% of their DEGs. Thus, each sequence has now an associated description and one or more GO ids. The other 18% of sequences remain unannotated and, consequently, cannot proceed with the functional analysis.

3.3.2 Functional Analysis:

We seek for overrepresented GO terms in our DEGs annotated pool by using the Fisher's exact test. Those GO terms which have an adjusted p-value lower than 0.05 are considered as overrepresented. Hence, we run the test for our three comparisons gene pool and using an annotated genome as reference. Results are depicted in figure 7.

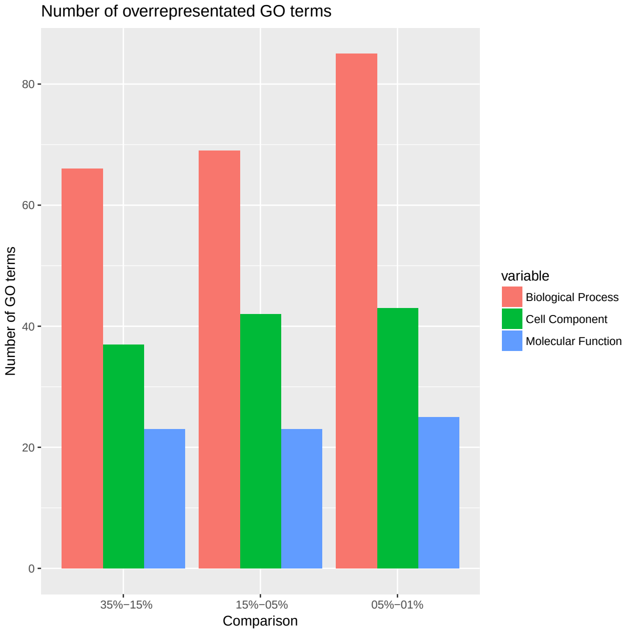

Figure 7: Number of overrepresentated GO terms obtained in each comparison.

So, we observe that 05%-01% is the comparison which has more overrepresented GO terms, which is linked to the fact that it also has the biggest DEGs pool. We also observe that biological process is the ontology with more hits for all comparisons. It is a more general ontology than other two so it makes sense that it capitalizes more hits. Moreover,

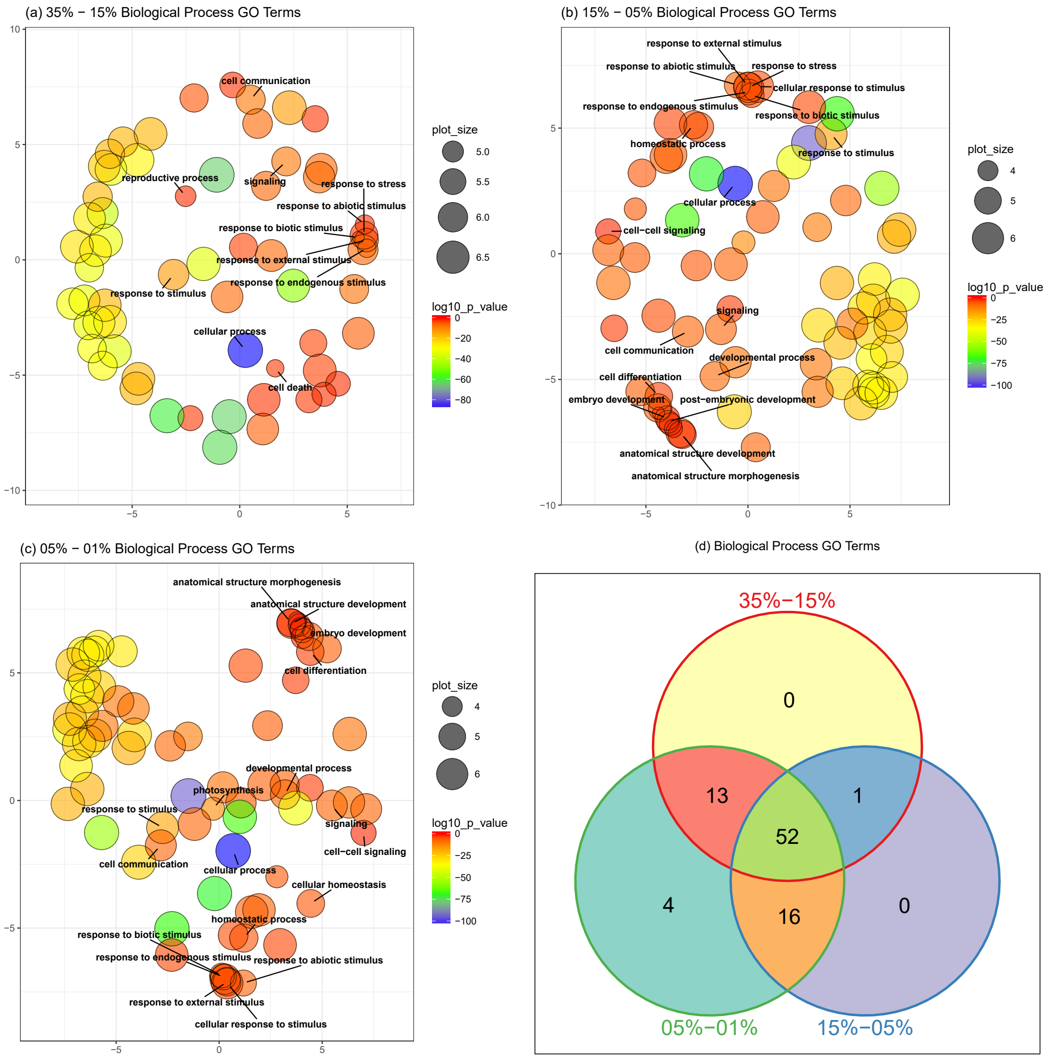

Figure 8: Biological process GO terms.(a) 35% - 15% comparison reviGO plot. (b) 15% - 05% comparison reviGO plot. (c) 05% - 01% comparison reivGO plot. (d) Leaf samples GO terms venn diagram.

we are seeking if differences in the expression pattern are linked to drought stress. Consequently, we decide to focus on biological process ontology rather than cell component or molecular function.

Next step in our enrichment analysis, is to run the Revigo clustering algorithm for the three comparisons. We are interested in the way the algorithm clusters and depicts GO terms rather than reducing our list. This way, we allow a large dispensability value and only exclude those GO terms from our list whose number is above 90%. Then we draw a Revigo plot for each comparison. We also draw a Venn diagram showing which biological process GO terms are common among comparisons. They are all together in figure 8. Not all GO terms are shown in the plot in order to allow a correct plot visualization. We observe that in all three comparisons, the most significant GO term (blue dot) is cellular process. It is followed by some green and mostly unrelated dots. Then, we find a cluster of yellow dots which are related to metabolic processes. Finally, we find the red dot clusters; although they are the ones with the lower log10 p-value, they are still significant and yield the highest relevance for our study.

There is a common cluster in all three comparisons which is formed by "response" related GO terms. Among others, we find the GO terms "response to stress", "response to external stimulus" and "response to abiotic stimulus" which can be all directly linked to drought stress. This way, among other causes, our expression pattern change is related to SWC conditions. Also, the GO term "homeostatic process", found in 15%-05% and 05%-01% comparisons, becomes highly relevant when talking of drought stress. As the SWC drops, the soil salinity increases involving both osmotic and ion toxicity effects on cells affecting its homeostasis [47].

There is also a cluster formed by "cell differentiation" related GO-terms, which is shared by the 15%-05% and 05%-01%

comparisons. Plants, as sessile organisms, have evolved to adapt and respond to several environmental stresses including drought conditions. Thus, growth and development is adjusted as a response to drought stress [28] involving the change in expression pattern that we detect.

Finally, and focusing on the Venn diagram, we observe that most of the GO terms are shared among the comparisons. In fact, only the last comparison (05%-01%) has 4 GO terms which are not found in the other comparisons. Among those 4 there is the GO term "photosynthesis"; drought stress has been related to a decrease in leaf water potential and stomatal opening, leading to a dysregulation of those genes related to photosynthesis [27]. Thus, we obtain another GO term related to drought pressure when it is at its highest rate. This reinforces the argument that the expression pattern change is led by the drought pressure.

Overall, the functional analysis allows us to stablish a direct link between the SWC conditions fixed in our experimental design and the detected DEGs. It is a consequence of the connection between the above-mentioned overepresentated GO terms and plant drought stress response literature. Those GO terms are linked to DEGs, which automatically become candidates for further analysis of *Lolium P.* drought stress tolerance. But we must remember that those DEGs are yet to be verified. Even when the scoring function is build around a validated data set and statistical evidence points out that a gene is a DEG a validation procedure needs to be performed, e.g. qPCR validation. Otherwise, we cannot assure that the expression pattern is truly being modified by our experimental conditons. Thus, this approach, along with other several RNA-seq pipelines, allows us to narrow and focus our point of view. But, given the magnitude of an average transcriptome, to end up with list of possible DEGs to focus on is a great result; and, actually, the aim of most RNA-seq studies.

4. CONCLUSIONS

Our goal is to build a scoring function to maximize and strengthen the detection of DEGs. Then, to test it with an experimental case which is the evaluation of *Lolium P.* expression pattern under drought stress conditions. When assessing for differentially expression using the four abovementioned packages we obtain different results, both for the validated and the experimental data sets. This highlight that RNA-seq analysis requires improvement. The scoring function, using a z-score mean approach, allows us to detect not only almost all the qPCR verified DEGs from the validated data set but include few false positive genes. Then, when applied to the experimental data set, it allows us to detect DEGs and link them the drought conditions. Consequently, our approach has been proven successful for detecting DEGs related to experimental conditions.

Even when the method works, there is room for its improvement. First, the qPCR validation gene pool was too little in respect from those considered as DEGs. For most of those genes considered as DEGs after the scoring function is applied, we do not know whether they are true or false positives. Thus, using a larger validated data set could strengthen our procedure allowing a better development of the scoring function. The function could also be improved by calibrating the weight of each package to the score; we compute the score using data from all the testing packages when we could use combinations of testing packages. Thus, it would allow us to determine which packages contribute the most to the score used for detecting DEGs or, on the contrary, which undermine the score. The procedure could also be improved with the addition of more software packages for DEGs testing.

Finally, we have a pool of DEGs related to drought response in *Lolium P.* leaf tissue. Thus, after the functional analysis of the remaining samples and once they are validated (e.g. via qPCR) we will have some target genes. Thus, further analysis based on those genes can be developed allowing us to characterize the molecular pathway behind *Lolium P.* drought stress response.

5. SUPPLEMENTARY MATERIAL

In this section, one can find the relation of supplementary material linked to the study

- **Supplementary File 1:** Generating count matrices
- **Supplementary File 2:** Pre-processing analysis
- **Supplementary File 3:** baySeq analysis
- **Supplementary File 4:** DESeq2 analysis
- **Supplementary File 5:** edgeR analysis
- **Supplementary File 6:** limma - voom analysis
- **Supplementary File 7:** Scoring function analysis
- **Supplementary File 8:** Validated data set differential expression testing results
- **Supplementary File 9:** Leaf Samples (against reference) differential expression testing results
- **Supplementary File 10:** Leaf Samples (time course) differential expression testing results
- **Supplementary File 11:** Root Samples (against reference) differential expression testing results
- **Supplementary File 12:** Root Samples (time course) differential expression testing results
- **Supplementary File 13:** Leaf Samples (time course) annotation results
- **Supplementary File 14:** Leaf Samples (time course) Fisher's test results
- **Supplementary File 15:** Leaf Samples (time course) reviGO results

References

[1] Altschul, S.F. et al. 1990. Basic local alignment search tool. *Journal of molecular biology.* 215, 3 (1990), 403–10.

[2] Anders, S. and Huber, W. 2010. Differential expression analysis for sequence count data. *Genome biology.* 11, (2010), R106.

[3] Anders, S. et al. 2012. Detecting diferential usage of exons from RNA-seq data. *Genome Research.* 22, 10 (2012), 2008–2017.

[4] Auer, P.L. and Doerge, R.W. 2011. A Two-Stage Poisson Model for Testing RNA-Seq Data. *Statistical Applications in Genetics and Molecular Biology.* 10, 1 (2011), 1–26.

[5] Belle, G. van et al. 2004. 13.4 ESTIMATING AND

SUMMARIZING ACCURACY. *Biostatistics: A methodology for the health sciences.* John Wiley & Sons. 558–560.

[6] Belle, G. van et al. 2004. 14.3 PRINCIPAL COMPONENTS. *Biostatistics: A methodology for the health sciences.* John Wiley & Sons. 588–599.

[7] Belle, G. van et al. 2004. 4.4 NORMAL DISTRIBUTIONS. *Biostatistics: A methodology for the health sciences.* John Wiley & Sons. 72–82.

[8] Benjamini, Y. and Hochberg, Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing.

[9] Blake, J.A. et al. 2015. Gene ontology consortium: Going forward. *Nucleic Acids Research.* 43, D1 (2015), D1049–D1056.

[10] Bourgon, R. et al. 2010. Independent filtering increases detection power for high-throughput experiments. *Pnas.* 107, 21 (2010), 9546–9551.

[11] Bullard, J.H. et al. 2010. Evaluation of Statistical Methods for Normalization and Differential Expression in mRNA-Seq Experiments Evaluation of Statistical Methods for Normalization and Differential Expression in mRNA-Seq Experiments. *BMC Bioinformatics.* 11, 1 (2010), 94.

[12] Byrne, S.L. et al. 2015. A synteny-based draft genome sequence of the forage grass Lolium perenne. *Plant Journal.* 84, 4 (2015), 816–826.

[13] Cock, P.J.A. et al. 2009. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic Acids Research.* 38, 6 (2009), 1767–1771.

[14] Conesa, A. et al. 2005. Blast2GO: A universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics.* 21, 18 (2005), 3674–3676.

[15] Conesa, A. et al. 2016. A survey of best practices for RNA-seq data analysis. *Genome Biology.* 17, 1 (2016), 13–19.

[16] Cook, R.D. 1977. Detection of Influential Observation in Linear Regression. *Technometrics.* 19, 1 (1977), 15–18.

[17] Finn, R.D. et al. 2017. InterPro in 2017-beyond protein family and domain annotations. *Nucleic Acids Research.* 45, D1 (2017), D190–D199.

[18] Grabherr, M.G. et al. 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology.* 29, 7 (2011), 644–652.

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

[20] Law, C.W. et al. 2014. voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biology.* 15, (2014), R29.

[21] Lawrence, M. et al. 2013. Software for Computing and Annotating Genomic Ranges. *PLoS Computational Biology.* 9, 8 (2013), 1–10.

[22] Li, B. and Dewey, C.N. 2011. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC bioinformatics.* 12, 1 (2011), 323.

[23] Li, H. et al. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics.* 25, 16 (2009), 2078–2079.

[24] Marioni, J.C. et al. 2008. RNA-seq : An assessment of technical reproducibility and comparison with gene expression arrays. *Genome Research.* 18, (2008), 1509–1517.

[25] Mortazavi, A. et al. 2008. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods.* 5, 7 (2008), 621–628.

[26] National Human Genome Research Institute (NHGRI) 2016. The Cost of Sequencing a Human Genome.

[27] Osakabe, K. and Osakabe, Y. 2012. Plant light stress. *Encyclopaedia of life sciences.* N.P. Group, ed.

[28] Osakabe, Y. et al. 2014. Response of plants to water stress. *Frontiers in Plant Science.* 5, March (2014), 86.

[29] Oshlack, A. et al. 2010. From RNA-seq reads to differential expression results. *Genome Biology.* 11, (2010), 220.

[30] R Development Core Team 2014. R: A language and evironment for statistical computing.

[31] Rajkumar, A.P. et al. 2015. Experimental validation of methods for differential gene expression analysis and sample pooling in RNA-seq. *BMC genomics.* 16, 1 (2015), 548.

[32] RNA-Seq Blog 2017. Number of publications citing RNA-Seq continues to increase.

[33] Robinson, M.D. and Oshlack, A. 2010. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biology.* 11, (2010), R25.

[34] Robinson, M.D. and Smyth, G.K. 2007. Moderated statistical tests for assessing differences in tag abundance. *Bionformatics Original Paper.* 23, 21 (2007), 2881–2887.

[35] Robinson, M.D. et al. 2009. edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics Original Paper.* 26, 1 (2009), 139–140.

[36] Schulz, M.H. et al. 2012. Oases: robust de novo RNA-seq assembly across the dynamic range of expression. *Bionformatics Original Paper.* 28, 8 (2012), 1086–109210.

[37] Seyednasrollah, F. et al. 2013. Comparison of software

packages for detecting differential expression in RNA-seq studies. *Briefings in Bioinformatics*. 16, 1 (2013), 59–70.

[38] Smyth, G.K. 2004. Statistical Applications in Genetics and Molecular Biology Linear Models and Empirical Bayes Methods for Assessing Differential Expression in Microarray. *Statistical Applications in Genetics and Molecular Biology*. 3, 1 (2004).

[39] Soneson, C. and Delorenzi, M. 2013. A comparison of methods for differential expression analysis of RNA-seq data. *BMC Bioinformatics*. 14, (2013), 91.

[40] Sultan, M. et al. 2008. A Global View of Gene Activity and Alternative Splicing by Deep Sequencing of the Human Transcriptome. *Science*. 685, August (2008), 956–960.

[41] Supek, F. et al. 2011. Revigo summarizes and visualizes long lists of gene ontology terms. *PLoS ONE*. 6, 7 (2011).

[42] Tang, F. et al. 2009. mRNA-Seq whole-transcriptome analysis of a single cell. *Nature Publishing Group*. 6, 5 (2009), 377–384.

[43] Trapnell, C. et al. 2011. Computational methods for transcriptome annotation and quantification using RNA-seq. *Nature Publishing Group*. 8, 6 (2011), 469–477.

[44] Trapnell, C. et al. 2013. Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nature biotechnology*. 31, 1 (2013), 46–53.

[45] Upton, G.J.G. 1992. Fisher's Exact Test. *Journal of the Royal Statistical Society*. 155, 3 (1992), 395–402.

[46] Zhang, Z.H. et al. 2014. A comparative study of techniques for differential expression analysis on RNA-seq data. *PLoS ONE*. 9, 8 (2014).

[47] Zhu, J.-K. 2016. Abiotic Stress Signaling and Responses in Plants. *Cell*. 167, 2 (2016), 313–324.