A hypothesis-driven approach to assessing significance of differences in RNA expression levels among specific groups of genes

Mingze He[1,2], Peng Liu[1,3], and Carolyn J. Lawrence-Dill[1,2,4*]
1. Bioinformatics and Computational Biology Program, Iowa State University, Ames, Iowa, USA, 50011
2. Department of Genetics, Development and Cell Biology, Iowa State University, Ames, Iowa, USA 50011
3. Department of Statistics, Iowa State University, Ames, Iowa, USA 50011
4. Department of Agronomy, Iowa State University, Ames, Iowa, USA 50011
*communications: triffid@iastate.edu (ORCID 0000-0003-0069-1430)

Abstract
Genome-wide molecular gene expression studies generally compare expression values for each gene across multiple conditions followed by cluster and gene set enrichment analysis to determine whether differentially expressed genes are enriched in specific biochemical pathways, cellular components, biological processes, and/or molecular functions, etc. This approach to analyzing differences in gene expression enables discovery of gene function, but is not useful to determine whether pre-defined *groups* of genes share or diverge in their expression patterns in response to treatments nor to assess the correctness of pre-defined gene set groupings. Here we present a simple method that changes the dimension of comparison by treating genes as variable traits to directly assess significance of differences in expression levels among pre-defined gene groups. Because expression distributions are typically skewed (thus unfit for direct assessment using Gaussian statistical methods) our method involves transforming expression data to approximate a normal distribution followed by dividing the genes into groups, then applying Gaussian parametric methods to assess significance of observed differences. This method enables the assessment of differences in gene expression distributions within and across samples, enabling hypothesis-based comparison among groups of genes. We demonstrate this method by assessing the significance of specific gene groups' differential response to heat stress conditions in maize.

Keywords
gene expression, statistical methods, RNA-seq, genomics

Abbreviations
GO – gene ontology
HSP – heat shock protein
KEGG – Kyoto Encyclopedia of Genes and Genomes
HSF TF – heat shock factor transcription factor
HSBP – heat shock binding protein
RNA – ribonucleic acid
TE – transposable element
TF – transcription factor
TPM – transcripts per kilobase millions

1. Introduction

Determining gene function remains a fundamental problem in biology. Measuring gene expression levels via transcript analysis across various treatments and developmental stages from many tissues greatly facilitates gene, pathway, and genomic functional annotation and interpretation. Because gene expression measurements are highly variable, many existing sophisticated statistical models and implementations are available to reduce measurement bias introduced during sampling and technical procedures[1,2,3,4,5,6,7]. After these calibration and normalization procedures, researchers commonly work toward identifying a set of differentially expressed genes then analyze annotations to those genes in an effort to discover shared functions (based on, e.g., shared biochemical pathways or biological processes; see Figure 1). Existing methods[8,9,10] for such analyses are highly dependent on the accuracy and precision of gene function annotations from gene ontology (GO) terms[11], KEGG pathway enzyme annotation[12], etc.[13] In other words, enrichment analysis depends on reliable functional annotation. Because some annotations are based on, e.g., experimental evidence whereas others derive from computational prediction alone,[13] annotation bias, incorrect annotations, and low quality annotations could lead to misinterpretation of detected expression patterns.

Other, less commonly used approaches assess differential expression within pre-defined gene sets[14,15]. Such approaches are either 'competitive' or 'self-contained,' terms coined by Goeman and Buhlmann[2]. The competitive approach identifies a gene set with comparatively greater or comparatively fewer differentially expressed genes relative to all other genes in the dataset. The null hypothesis for the competitive approach is as follows: "Differential expression in a defined subset of genes is observed, at most, as often as it is observed for genes in the remaining set." The self-contained approach focuses only on the information from gene sets of interest. The null hypothesis underlying self-contained tests could be stated as, "No genes in the defined subset are differentially expressed." Each approach has important caveats. Competitive group analysis depends on the background distribution and assumes independent sampling[2,16,17]. Self-contained analyses can be highly affected by a very small number of highly expressed genes; thus one highly expressed gene could result in failure to detect otherwise significant patterns. Another drawback to these methods is the fact that enrichment analyses, designed to test whether specific gene groups are either up- or down-regulated under specific conditions enriched in certain biological functions, do not produce a direct measurement of how *much* gene group expression levels change across experiments.

Here we present and describe a different method to comparatively assess the expression patterns among pre-defined groups, both within and across treatments. The key difference between existing methods (Figure 1) and ours (Figure 2) is the absence of a 'differentially expressed gene set' concept throughout the process. We treat each normalized gene expression value as a variable in the later comparison between groups. Our method enables examination on non-linear, non-one-to-one correspondence among gene groups.

We demonstrate this new method by assessing a well-studied biological phenomenon: the response of genes to heat stress. Heat shock proteins (HSP) are regulated by heat shock factors (which are a specific group of transcription factors; abbreviated here HSF TFs) under heat stress[18,19]. HSF TFs are negatively regulated by heat shock factor protein bindings (HSPBs)[20]. By dividing maize genes into subgroups, i.e., HSPs, HSPBs, HSFs, other TFs, and housekeeping genes, we compare each group's response to heat stress.

We also show the method's utility for analyzing groups of genes that are not functionally related by evaluating the significance of differences in gene expression for maize genes nearby *gyma* and *huck* transposable elements (TE) in heat stress conditions, confirming Makarevitch et al.'s findings that genes with *gyma* nearby are significantly up-regulated whereas genes with *huck* nearby show no differential expression in heat stress conditions[21].

2. Materials and Methods

For the first example, we focused on these subgroups of maize genes: HSPs (reported in Pegoraro *et al.*[22]), HSPBs (from Gramene[23,24]), HSF TFs and other TFs (from GRASSIUS[25,26]), and housekeeping genes (see "Supplementary Material 2" from Lin *et al.*[11]). Each group's response to heat stress was compared across conditions and across subgroups within the same

condition using RNA-seq datasets reported by Makarevitch *et al.*[21] Gene groups derived from these sources are listed in Supplementary Table S1 (all Supplementary Materials are available online at DOI 10.17605/OSF.IO/Y46TF, which currently resolves to https://osf.io/y46tf/).

For the second example, we use the Makarevitch *et al.*[21] expression dataset along with their TE gene set lists and coordinates from the maize TE dataset reported by Baucom *et al.* [27] where TEs are considered 'nearby' if they are +/-1 kb from the transcription start sites of genes in the B73 RefGen_v3 genome assembly with gene annotation version ZmB73_5b[28]. We focus on genes near *gyma* TEs, which constituted the largest set of genes near a specific TE type that were up-regulated by heat stress, and compare those to genes near *huck*, which constituted the largest subset of genes near a TE family that were not differentially expressed. Gene groups derived from these sources are listed in Supplementary Table S2.

Because this method aims to detect the differences between groups of genes based on a hypothesis-driven approach, methods for grouping genes are not limited to genes with homologous sequences (such as groups in the first example) or shared pathways, though gene groups assembled using such concepts are reasonable candidates for this approach. One example of a non-homologous gene set made up of members that also are not necessarily involved in shared biochemical pathways includes members of the housekeeping gene set (such as those used in the normalization step described in Theory/Calculations), which we define as those genes essential to maintaining basic cellular functions and including only those with constant expression values across various tissues and conditions. Other groups of genes share such characteristics as having a TE located near the transcription start site (the second example), genes in a particular region of a chromosome, or genes sharing structural elements. These groups each share some characteristic, but these characteristics are not limited.

3. Theory/Calculations

Suppose we have a pre-defined gene set, for which read counts for each gene have been normalized by TPM that is calculated as a percentage of reads normalized by length of each gene to total reads within a sample[29]. Because we intend to treat each gene as a variable, we use the TPM value averaged across biological replicates to measure the expression level for each gene:

$$G_{gi}=\frac{1}{n}\sum_{j=1}^{n} T_{gij}, \qquad (1.0)$$

where T_{gij} stands for the TPM value of the gth gene for the jth biological replicates within the ith condition/treatment, j = 1,2, …, n_i, and n_i is the number of replicates for the ith condition. This step summarizes expression levels for each gene to facilitate the comparison using gene sets.

We apply the logarithm transformation as in (1.1) because the expression measurements are typically highly skewed to the right. We add a small value 1 to each TPM to avoid problems for the logarithm transformation.

$$X_{gi}= \log (G_{gi} + 1) \qquad (1.1)$$

Secondly, we calculate a normalization factor based on housekeeping genes:

$$m_i =\frac{1}{r}\sum_{g=1}^{r} HX_{gi} \qquad (1.2)$$

where HX_{gj} are expression values of the gth housekeeping gene in the ith condition after transformation using (1.0) and (1.1), and suppose that there are a total of r housekeeping genes.

Because housekeeping genes are, by definition, expected to express at a constant level across tissues and under various conditions, we could use m_i to normalize across conditions (X_{gi}/m_i) when we compare a gene set from different samples.

Finally, Student's *t*-test (or other appropriate parametric or non-parametric tests) could be applied to test two different types of null hypotheses.

First, we can test for equal means between gene groups A and B under the same treatment condition, e.g., condition 1 (1.3).

$$H_0: \mu_{A1} = \mu_{B1}$$
$$H_a: \mu_{A1} \neq \mu_{B1} \qquad (1.3)$$

This might be of interest to determine, e.g., whether a transcription factor and the genes it activates respond similarly in a given condition.

Second, we could also test whether the same gene set (A) has the same mean across conditions 1 and 2:

$$H_0: \mu_{A1} = \mu_{A2}$$
$$H_a: \mu_{A1} \neq \mu_{A2} \qquad (1.4)$$

This might be of interest to determine, e.g., whether a group of genes with increased transcript abundance under salt stress conditions shows similar increases in transcript abundance in drought stress conditions.

In both examples, if a gene has an average TPM value (G_{gi}) below 1 in both conditions, we consider such data unreliable because it is hard to tell whether the corresponding reads were from signal or noise. Therefore we exclude such genes from the following analysis. Note that we used TPM in the above formulae, but FPKM values can also be used.

4. Results and Discussion

As is often observed for gene expression values, the distribution of expression values across all genes in shoot tissues from maize seedlings follows an exponential curve, with the majority of genes expressed at relatively low levels (Figure 3a; non-stressed condition). Log transformation results in a distribution much closer to normal (Figure 3b). Log-transformed data collected from the shoot tissues of maize seedlings under both non-stress (Figure 3c) and heat stress conditions (Figure 3d) generally follow the normal distribution (i.e., 93.1%-97.4% of all log transformed data were located within a 95% confidence interval), indicating that this transformation approach is reasonable. Note well: because our method relies upon transformation to approximate a normal distribution, it is important to check the results of log transformation not only for all sampled genes and for all conditions, but also for each individual gene group and treatment combination. In this example, the housekeeping genes, HSFs, other TFs, and HSPBs all appear to roughly approximate the normal distribution (as assessed via QQ-plot; Figs. S1-8). However, the transformed expression patterns for HSPs do not follow the normal distribution (see Figs. S9 and S10) and are therefore not appropriate for analyses using parametric tests of significance among groups. This result exemplifies the need to inspect transformed distributions as a step in applying this method.

As one might expect given the known and well-understood biology of response to heat stress, transcription of HSF TFs increases in response to heat stress and shows a very different distribution than other TFs (Figure 3 panels e and f). Relative to the non-stress condition, HSF TFs have right-shifted RNA expression distributions relative to housekeeping genes and other TFs under heat stress. Beyond inspecting the distributions, this data transformation approach allows application of parametric statistical approaches, e.g., the *t*-test, to compare mean values between distributions within a given sample. As shown in Table 1, under non-stress conditions, the *t*-test fails to reject the null hypothesis (i.e., HSF TF and other TF have no differences in mean values). However, as shown in Table 2, under heat stress *t*-test results reject the null hypothesis, indicating that the higher expression of HSF TFs is significantly different than that

of other TFs. As shown in Table 3, the expression distribution shifts between Figure 1 panels e and f are significant only for HSF TFs, but not for other TFs nor for the HSPB group.

Table 1. *t*-test p-values between gene sets under non-stress conditions.

	Sample size	HSF TF	HSPB	Other TF
HSF TF	19	-	<0.0004*	0.272
HSPB	37	-	-	0.0001*
Other TF	1,299	-	-	-

* p-values smaller than 0.05 after Bonferroni multiple test correction.

Table 2. *t*-test p-values between gene sets under stress conditions.

	Sample size	HSF TF	HSPB	Other TF
HSF TF	23	-	0.5003	0.0060*
HSPB	37	-	-	<0.0001*
Other TF	1,234	-	-	-

* p-values smaller than 0.05 after Bonferroni multiple test correction.

Table 3. *t*-test p-values among the same groups of genes under two conditions.

	HSF TF	HSPB	Other TF
Heat *vs* normal	0.0039*	0.1659	0.2457

* p-values smaller than 0.05 after Bonferroni multiple test correction.

It is worthwhile to show that the approach we have developed is versatile beyond the analysis of homologous genes with shared functions given that hypotheses concerning concerted action of genes can be made for various groups (e.g., genes that share specific TF binding sites, genes that all participate in the same biological pathway, etc.). For our second example, we investigated differences in expression patterns among genes near TE types that Makarevitch *et al*. showed may affect transcription under heat, cold, high UV, and salt stress conditions[21]. We analyzed their maize RNA-seq dataset collected under normal and heat stress using our new method as described in the first example. Among other findings, Makarevitch *et al.* showed that 25% of genes with a *gyma* element +/-1 kb of the transcription start site are up-regulated under heat stress whereas those with the *huck* element in the same location are not[21]. We plot transformed expression patterns of groups of genes with the *gyma* and *huck* motifs near the transcription start site (as well as housekeeping genes) under non-stress and heat stress scenarios (Figure 4). Those plots show that the distribution of genes with *gyma* near the transcription start site shifts to the right under heat stress conditions, which indicates a positive transcriptional response to heat stress. Genes with *huck* near the transcription start site do not. We then apply our method to assign p-values to assess the significance of both *gyma* and *huck* to the activation of nearby gene expression (for QQ-plots see Figs. S11-S14). The *gyma* and *huck* distribution means are significantly different from each other in non-stress conditions (Table 4), but in heat stress conditions the means of *gyma* and *huck* expression distributions are not significantly different from each other (Table 5). Compared across samples (non-stress *vs* heat stress), the expression differences for the *gyma* gene group means yields a p-value < 0.0001 (see Table 6). This indicates a statistically significant positive response to heat stress for genes with *gyma* nearby. Genes with the *huck* TE family nearby show no significant changes in the distribution of transformed expression values (p-value=0.76). These results not only confirm the findings of Makarevitch et al., they assign significance to the observation that genes with *gyma*, but not *huck*, near the transcription start site show increased transcription under heat stress[21].

Table 4. *t*-test p-values between gene sets under non-stress conditions.

	Sample size	*gyma*	*huck*
gyma	11	-	<0.0001*
huck	234	-	-

* p-values smaller than 0.05 after Bonferroni multiple test correction.

Table 5. *t*-test p-values between gene sets under stress conditions.

	Sample size	*gyma*	*huck*
gyma	11	-	0.4640
huck	213	-	-

Table 6. *t*-test p-values among the same groups of genes under two conditions.

	gyma	*huck*
Heat stress *vs* non-stress	<0.0001*	0.7575

* p-values smaller than 0.05 after Bonferroni multiple test correction.

5. Conclusions

We changed the dimension of comparison for gene expression studies to derive a new hypothesis-driven approach to gene expression analysis that enables the use of Gaussian parametric tests to assess results. One could also easily use this approach to study other phenomena and to test various biological hypotheses. For example, recent studies reporting gene regulation show that motifs around regulatory regions of genes (e.g., transposable elements[21], high GC content motifs[30], and G-quadruplexes[31]) may influence nearby gene expression levels under stress conditions. One could also apply this approach to evaluate the influence of gene sequence composition bias on expression (e.g., GC content effects) as well as expression differences that may be attributable to a gene's local context (e.g., location on the chromosome or adjacency to other genes).

By design, our approach is limited to pre-defined gene groups. Thus, this method is not suitable to discover genes that respond to, e.g., a particular stress environment, which is often the goal of a typical discovery-based approach to gene expression analysis. Also, our approach does not attempt to assess the correctness of gene groupings. For example, if some HSF TF had been missing from the HSF TF gene group, the analysis could still yield significant p-values. On the other hand, the power of this method is that it can be combined with existing knowledge about specific groups of genes to assess how well a given annotation method grouped genes based on shared gene expression patterns.

This new method is complementary to existing expression analysis tools such as those available via the Gramene Plant Reactome[32]. First, our method is not limited to homologous genes or pathway membership, thus, we broaden the scope of comparison that can be made. Second, our distribution-based visualization is a graph-based alternative to heatmap or color-coded plots used in pathway analysis, which can be difficult to interpret (especially for individuals with some types of colorblindness). Third, our method allows for statistical tests among groups of genes rather than between individual genes, where individual gene comparisons may conflict and cause confusing results for downstream analyses including, e.g., pathway analysis. Our test could be used to eliminate such a conflict by transforming each gene considered into a variable, and all genes in a certain pathway into a group.

Novel approaches to enable hypothesis testing for shared regulation of gene sets are needed. As technologies for measuring gene expression improve, we anticipate that additional

methods will be developed that expand the toolset for grouping-oriented gene expression analytics.

Acknowledgements

Funding in support of this work came from the Iowa State University Plant Sciences Institute (CJLD and MH) and NSF PGRP #1238142. We thank Irina Makarevitch at Hamline University for providing TE annotation files and for review. We thank Jennifer Clarke at University of Nebraska – Lincoln, Drena Dobbs and Justin Walley at Iowa State University, and members of the Lawrence-Dill Plant Informatics and Computation Lab for suggestions and helpful discussions.

References

1. Goeman, J. J., Oosting, J., Cleton-Jansen, A. M., Anninga, J. K. & van Houwelingen, H. C. Testing association of a pathway with survival using gene expression data. *Bioinformatics* **21,** 1950–1957 (2005).
2. Goeman, J. J. & Bühlmann, P. Analyzing gene expression data in terms of gene sets: Methodological issues. *Bioinformatics* **23,** 980–987 (2007).
3. Robinson, M. D. & Oshlack, A. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol.* **11,** R25 (2010).
4. Li, B. & Dewey, C. N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* **12,** 323 (2011).
5. Trapnell, C. *et al.* Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat. Protoc.* **7,** 562–78 (2012).
6. Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome Biol.* **11,** R106 (2010).
7. Patro, R., Duggal, G., Love, M. I., Irizarry, R. A. & Kingsford, C. Salmon provides fast and bias-aware quantification of transcript expression. *Nat. Methods* (2017).
8. Subramanian, A. *et al.* Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci.* **102,** 15545–15550 (2005).
9. Carbon, S. *et al.* AmiGO: Online access to ontology and annotation data. *Bioinformatics* **25,** 288–289 (2009).
10. Huang, D. W., Lempicki, R. a & Sherman, B. T. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.* **4,** 44–57 (2009).
11. Ashburner, M. *et al.* Gene ontology: Tool for the unification of biology. *Nature Genetics* **25,** 25–29 (2000).
12. Ogata, H. *et al.* KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Research* **27,** 29–34 (1999).
13. Rhee, S. Y., Wood, V., Dolinski, K. & Draghici, S. Use and misuse of the gene ontology annotations. *Nat. Rev. Genet.* **9,** 509–15 (2008).
14. Nam, D. & Kim, S. Y. Gene-set approach for expression pattern analysis. *Briefings in Bioinformatics* **9,** 189–197 (2008).
15. Guo, W., Yang, M., Xing, C. & Peddada, S. D. Analysis of high dimensional data using pre-defined set and subset information, with applications to genomic data. *BMC Bioinformatics* **13,** 177 (2012).
16. Allison, D. B., Cui, X., Page, G. P. & Sabripour, M. Microarray data analysis: from disarray to consolidation and consensus. *Nat Rev Genet* **7,** 55–65 (2006).
17. Damian, D. & Gorfine, M. Statistical concerns about the GSEA procedure. *Nat. Genet.* **36,** 663 (2004).
18. Sorger, P. K. Heat shock factor and the heat shock response. *Cell* **65,** 363–366 (1991).
19. Al-Whaibi, M. H. Plant heat-shock proteins: A mini review. *J. King Saud Univ. - Sci.* **23,** 139–150 (2011).
20. Fu, S., Rogowsky, P., Nover, L. & Scanlon, M. J. The maize heat shock factor-binding protein paralogs EMP2 and HSBP2 interact non-redundantly with specific heat shock factors. *Planta* **224,** 42–52 (2006).
21. Makarevitch, I. *et al.* Transposable Elements Contribute to Activation of Maize Genes in Response to Abiotic Stress. *PLoS Genet.* **11,** (2015).
22. Pegoraro, C., Mertz, L., da Maia, L., Rombaldi, C. & de Oliveira, A. Importance of heat shock proteins in maize. *J. Crop Sci. Biotechnol.* **14,** 85–95 (2011).
23. Tello-Ruiz, M. K. *et al.* Gramene 2016: Comparative plant genomics and pathway resources. *Nucleic Acids Res.* **44,** D1133–D1140 (2016).
24. Gramene at ftp://ftp.gramene.org/pub/gramene/CURRENT_RELEASE/data/ontology/go/go_ensembl_zea_mays.gaf term GO:0031072 (heat shock protein binding) accessed 3 May 2017.
25. Yilmaz, A. *et al.* GRASSIUS: a platform for comparative regulatory genomics across the grasses. *Plant Physiol.* **149,** 171–80 (2009).
26. GRASSIUS at http://grassius.org/tf_browsefamily.html?species=Maize accessed 3 May

2017.
27. Baucom, R. S. *et al.* Exceptional diversity, non-random distribution, and rapid evolution of retroelements in the B73 maize genome. *PLoS Genet.* **5,** (2009).
28. Schnable, P. S. *et al.* The B73 maize genome: complexity, diversity, and dynamics. *Science* **326,** 1112–5 (2009).
29. Wagner, G. P., Kin, K. & Lynch, V. J. Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. *Theory Biosci.* **131,** 281–285 (2012).
30. Mishra, A. K., Agarwal, S., Jain, C. K. & Rani, V. High GC content: critical parameter for predicting stress regulated miRNAs in Arabidopsis thaliana. *Bioinformation* **4,** 151–4 (2009).
31. Andorf, C. M. *et al.* G-Quadruplex (G4) Motifs in the Maize (Zea mays L.) Genome Are Enriched at Specific Locations in Thousands of Genes Coupled to Energy Status, Hypoxia, Low Sugar, and Nutrient Deprivation. *J. Genet. Genomics* **41,** 627–647 (2014).
32. Naithani, S. *et al.* Plant Reactome: A resource for plant pathways and comparative analysis. *Nucleic Acids Res.* **45,** D1029–D1039 (2017).

Figure Legends

Figure 1. Discovery-based approach to gene expression analysis. Boxes represent data or analysis results, arrows represent analyses. Step 1: RNA-seq samples are subjected to log fold change analysis to identify a DE gene set. Step 2: clustering algorithms are used to create DE clusters (groups 1, 2, and 3). Note that step 2 maybe skipped, especially when the list of DE genes is not long. Step 3: Enrichment analysis identifies pathways, GO terms, KEGG terms, etc. associated with the DE gene groups.

Figure 2. Our hypothesis-based approach to gene expression analysis. Boxes represent data or analysis results, arrows represent analyses. Step 1: RNA-seq samples (control and treatment) are subjected to raw read count normalization (transcripts per kilobase million; TPM) and log transformation for each gene. Step 2: Pre-defined gene groups (Control 1, Control 2, Control 3, Treatment a1, Treatment a2, etc. where the gene list for C1 is the same as Ta1 and Tb1, etc.) are defined and housekeeping genes are designated for normalization and to create group estimated means. Step 3: Gaussian statistical tests are carried out within (e.g., μ_{Ta1} and μ_{Ta2}) and/or across samples (e.g., μ_{Tb2} and μ_{Tb2}).

Figure 3. Log transformation enables Gaussian modeling of expression patterns among groups of genes. **(a)** The density plot of all maize genes (N=39,657) with a given RNA expression level is plotted for the non-stress condition. **(b)** Log transformation of RNA expression values before m normalization results in a roughly normal distribution. (Note y axis is not the same between panels a and b.) A QQ-plot (normal distribution quantiles plotted against sample quantiles) for the log-transformed data collected from **(c)** non-stress and **(d)** heat stress shown as black circles. Red diagonal indicates perfect concordance. Red dashed lines indicate the 95% confidence interval (CI). Percentage of data falling outside the 95% CI are indicated in purple. **(e** and **f)** RNA expression levels of HSF TF, HSPB and other TF gene groups normalized by housekeeping genes are plotted by percentage. Housekeeping genes shown in green, HSPB in red, HSF TFs in turquoise, and other TF family genes in purple. **(e)** Non-stress condition. **(f)** Heat stress.

Figure 4. TE within and near the promoter region activate nearby gene transcriptional response to heat stress. RNA expression levels of gene groups with *gyma* or *huck* near the promoter region normalized by housekeeping genes are plotted by percentage. Housekeeping genes shown in green, *gyma* in blue, and *huck* in red. **(a)** Non-stress condition. **(b)** Heat stress.

Figures

Figure 1

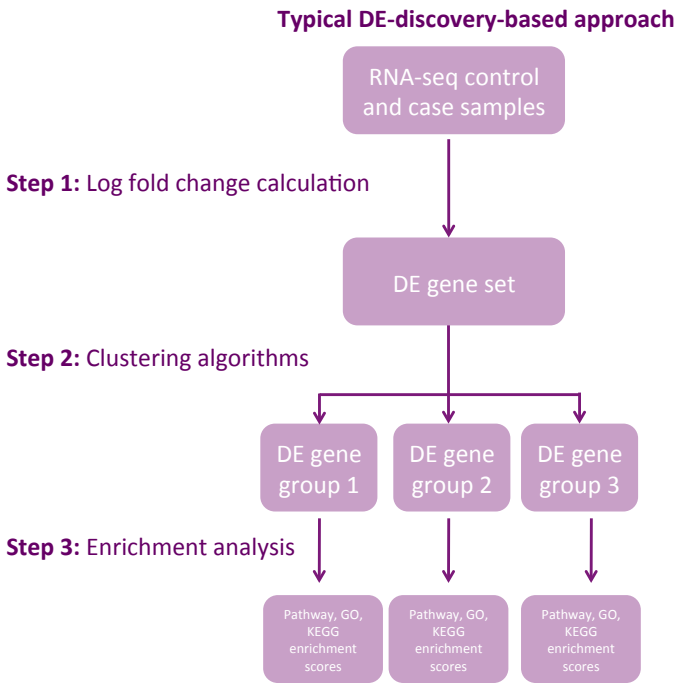


Figure 2

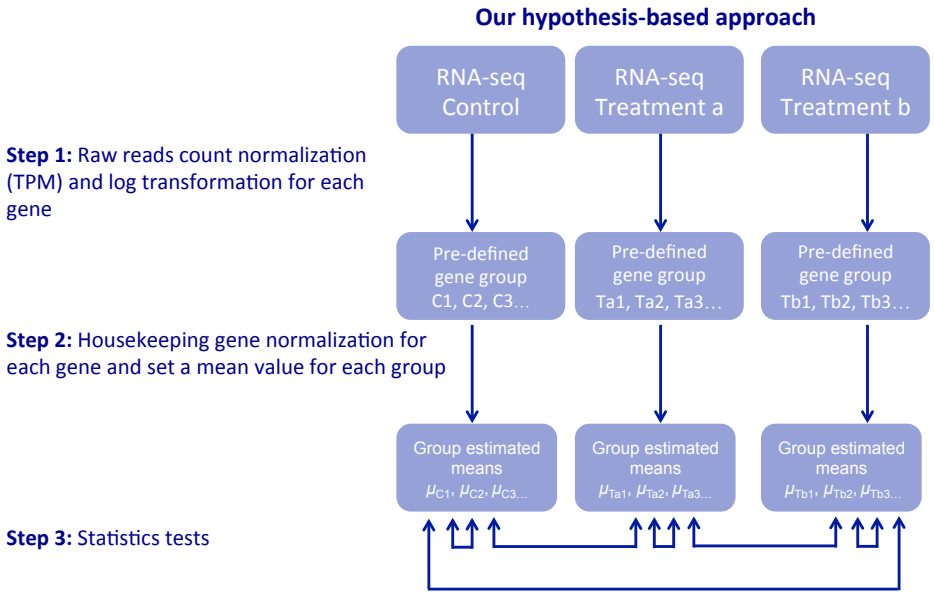

Our hypothesis-based approach
RNA-seq Control
RNA-seq Treatment a
RNA-seq Treatment b
Step 1: Raw reads count normalization (TPM) and log transformation for each gene
Pre-defined gene group C1, C2, C3…
Pre-defined gene group Ta1, Ta2, Ta3…
Pre-defined gene group Tb1, Tb2, Tb3…
Step 2: Housekeeping gene normalization for each gene and set a mean value for each group
Group estimated means μ_{C1}, μ_{C2}, μ_{C3}…
Group estimated means μ_{Ta1}, μ_{Ta2}, μ_{Ta3}…
Group estimated means μ_{Tb1}, μ_{Tb2}, μ_{Tb3}…
Step 3: Statistics tests

Figure 3

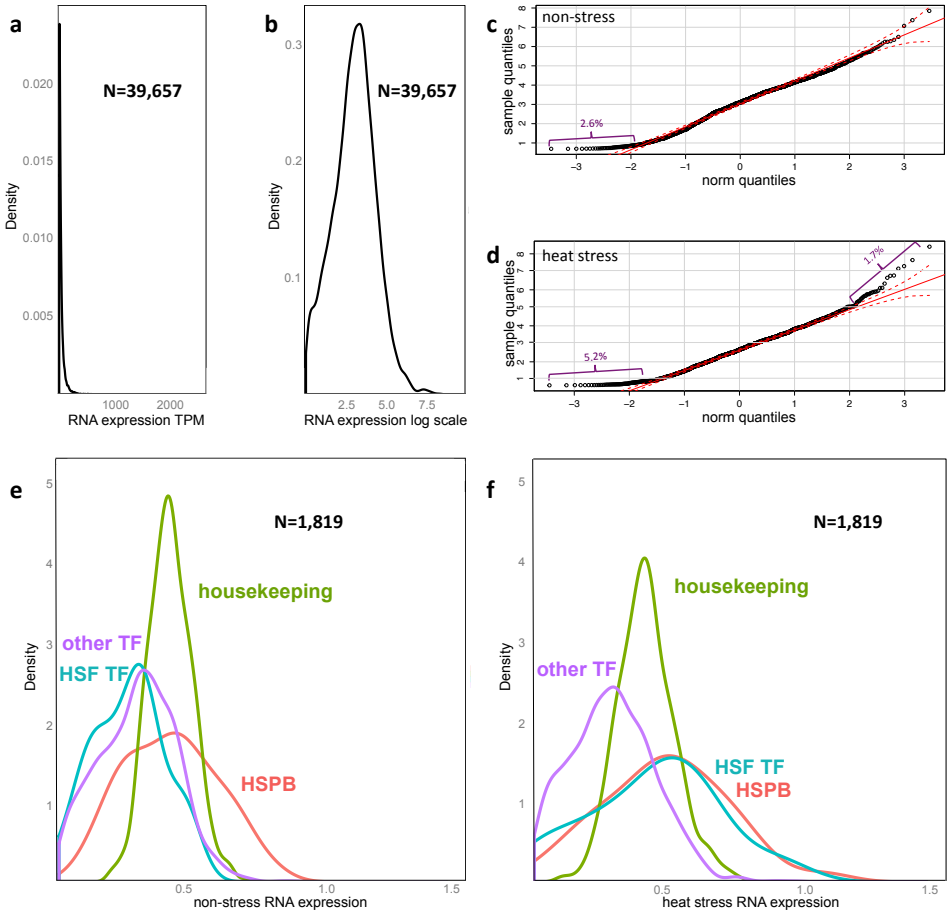
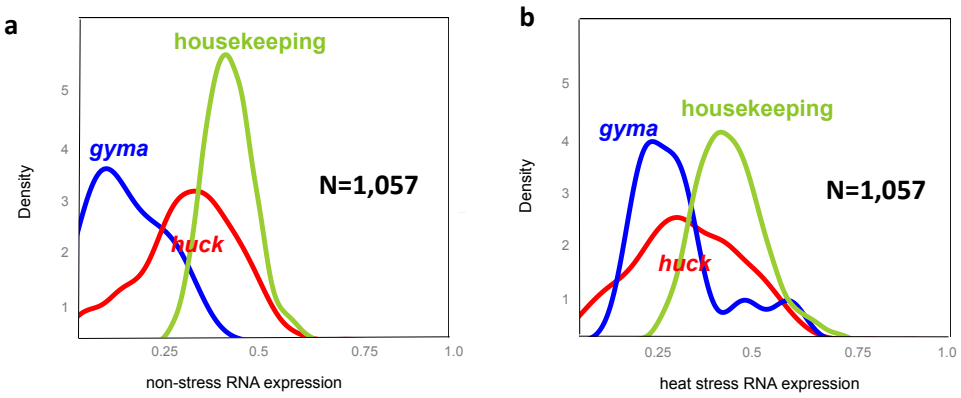

a
N=39,657
Density
RNA expression TPM
b
N=39,657
Density
RNA expression log scale
c
non-stress
2.6%
sample quantiles
norm quantiles
d
heat stress
1.7%
5.2%
sample quantiles
norm quantiles
e
N=1,819
housekeeping
other TF
HSF TF
HSPB
Density
non-stress RNA expression
f
N=1,819
housekeeping
other TF
HSF TF
HSPB
Density
heat stress RNA expression

Figure 4


a
housekeeping
gyma
huck
N=1,057
Density
5
4
3
2
1
0.25
0.5
0.75
1.0
non-stress RNA expression
b
housekeeping
gyma
huck
N=1,057
Density
5
4
3
2
1
0.25
0.5
0.75
1.0
heat stress RNA expression