A method to assess 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

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 and its caveats by assessing the significance of specific gene groups' differential response to heat stress conditions in maize.

Determining gene function remains a fundamental problem in biology. Measuring gene expression levels via RNA-seq analyses across various treatments and developmental stages from many tissues greatly facilitates gene, pathway, and genomic functional annotation and interpretation. Many sophisticated statistical models and implementations have been developed to reduce measurement bias introduced during sampling and technical procedures[1,2,3,4,5]. Following normalization, downstream analyses commonly aim to discover the function of genes that share differential expression (DE) patterns based on, e.g., shared biochemical pathways, biological processes, etc. Other, less commonly used approaches assess DE within pre-defined gene sets[6,7]. Such existing approaches are either 'competitive' or 'self-contained,' terms coined by Geoman and Buhlmann[8]. The competitive approach identifies gene sets enriched with more or less DE as compared to the background gene set. The self-contained approach focuses only on the information from gene sets of interest. Each has important caveats. Competitive group analysis depends on the background distribution and assumes independent sampling[8,9,10]. Self-contained analyses are highly affected by extreme values of expression for single genes, thus one highly expressed gene could result in failure to detect otherwise significant patterns. Here we present a different method: expression data transformation followed by Gaussian statistical assessment that enables comparative assessment of expression patterns among pre-defined groups, both within and across treatments.

After read count normalization (e.g., transcript per million [TPM]), which takes various transcript lengths into consideration, and normalization by, e.g., housekeeping gene mean values, transcription rates can be compared among individual genes or groups of genes, both within and across samples. In such situations, the distribution of expression values across all genes generally follows an exponential curve, with the majority of genes expressed at relatively low levels (e.g., see Figure 1a). Through log transformation, these

distributions become approximately normal (Figure 1b), thus enabling downstream analysis of differences among specific groups of genes including parametric approaches (e.g., Student *t*-test) to determine the significance of differences among groups based on expression pattern differences. To demonstrate this approach, we use a well-known phenomenon: the response of genes to heat stress. Heat shock proteins (HSP) are regulated by heat shock factors (HSFs, a type of transcription factor or TF) under heat stress[11,12]. HSPs bind specific genes called heat shock protein binding (HSPB) genes[13,14]. By dividing maize genes into subgroups, i.e., HSPs (reported in Pegoraro *et al.*[15]), HSBPs (from Gramene[13,14]), HSFs and other TFs (from GRASSIUS[16,17]), and housekeeping genes (see "Supplementary Material 2" from Lin *et al.*[11]), we compare each group's response to heat stress using RNA-seq datasets reported by Makarevitch *et al.*[18].

As expected, the expression pattern of shoot tissues of maize seedlings is extremely positively skewed (Figure 1a; non-stressed condition). Log transformation results in a distribution much closer to normal (Figure 1b). Log-transformed data collected from the shoot tissues of maize seedlings under both non-stress (Figure 1c) and heat stress conditions (Figure 1d) 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. 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 combinations. In this example, the housekeeping genes, HSFs, other TFs, and HSBPs all appear to roughly approximate the normal distribution (as assessed via QQ-plot; Figs. S1-8). However, the transformed expression pattern for HSPs do not follow the normal distribution (see Figs. S9 and S10) and are thus not appropriate for downstream parametric tests of significance among groups. This result exemplifies the need to inspect transformed distributions as a step in applying this method.

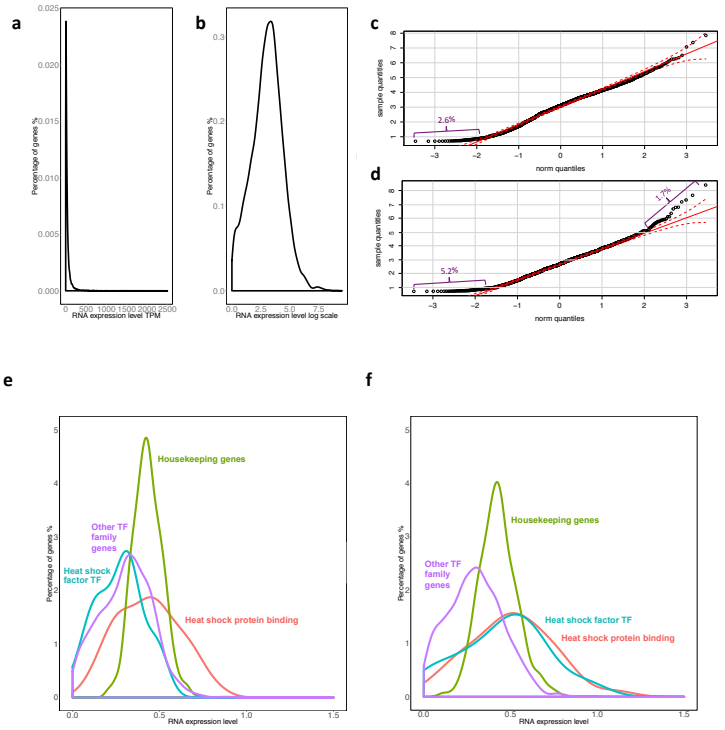

Figure 1 (previous page). Log transformation enables Gaussian modeling of expression patterns among groups of genes. **(a)** The percentage of maize genes with a given RNA expression level is plotted for the non-stress condition. **(b)** Log transformation of RNA expression values results in a roughly 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 normalized by housekeeping genes are plotted by percentage. Housekeeping genes shown in green, HSPB in red, HSF TFs in green, and other TF family genes in purple. **(e)** Non-stress condition. **(f)** Heat stress.

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 1 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 normal condition *t*-test fails to reject null hypothesis that HSF TF and other TF have different 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 HSBP 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 among the same groups of genes 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.

One could 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 (such as transposable elements[12], high GC content motifs[13], and G-quadruplex[14]) may influence nearby gene expression levels under

stress conditions. We 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). Lastly, this method could reassess the reasonableness of pre-defined gene sets based on homology or other sequence features.

Novel approaches to enable hypothesis testing for shared regulation of gene sets are needed. The approach we developed and report here is anticipated to be among the first of many such grouping-oriented analytics approaches under development.

Acknowledgements
Funding in support of this work came from the Iowa State University Plant Sciences Institute (CJLD and MH). We thank Jennifer Clarke at University of Nebraska – Lincoln for critical review and suggestions.

References

1. Robinson, M. D. & Oshlack, A. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol.* **11,** R25 (2010).
2. Li, B. & Dewey, C. N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* **12,** 323 (2011).
3. Trapnell, C. *et al.* Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat. Protoc.* **7,** 562–78 (2012).
4. Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome Biol.* **11,** R106 (2010).
5. 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).
6. Nam, D. & Kim, S. Y. Gene-set approach for expression pattern analysis. *Briefings in Bioinformatics* **9,** 189–197 (2008).
7. 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).
8. Goeman, J. J. & Bühlmann, P. Analyzing gene expression data in terms of gene sets: Methodological issues. *Bioinformatics* **23,** 980–987 (2007).
9. 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).
10. Damian, D. & Gorfine, M. Statistical concerns about the GSEA procedure. *Nat. Genet.* **36,** 663 (2004).
11. Sorger, P. K. Heat shock factor and the heat shock response. *Cell* **65,** 363–366 (1991).
12. Al-Whaibi, M. H. Plant heat-shock proteins: A mini review. *J. King Saud Univ. - Sci.* **23,** 139–150 (2011).
13. Tello-Ruiz, M. K. *et al.* Gramene 2016: Comparative plant genomics and pathway resources. *Nucleic Acids Res.* **44,** D1133–D1140 (2016).
14. 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.
15. 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).
16. Yilmaz, A. *et al.* GRASSIUS: a platform for comparative regulatory genomics across the grasses. *Plant Physiol.* **149,** 171–80 (2009).
17. GRASSIUS at http://grassius.org/tf_browsefamily.html?species=Maize accessed 3 May 2017.
18. Makarevitch, I. *et al.* Transposable Elements Contribute to Activation of Maize Genes in Response to Abiotic Stress. *PLoS Genet.* **11,** (2015).

Supplementary Figures

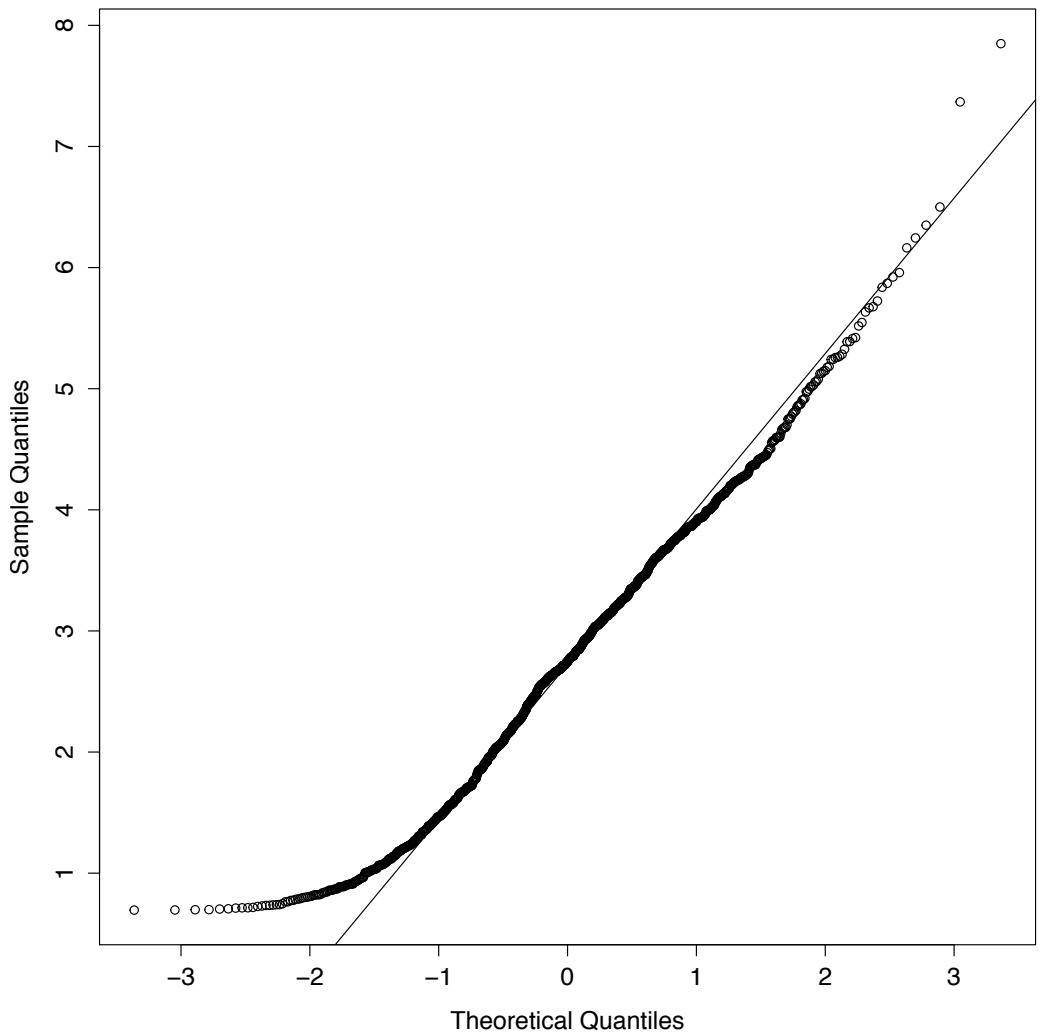

Figure S1. QQ-plot of housekeeping gene transcription levels under normal conditions. Normal distribution quantiles plotted against sample quantiles for the log-transformed data collected from non-stress shown as black circles. Black diagonal indicates perfect concordance.

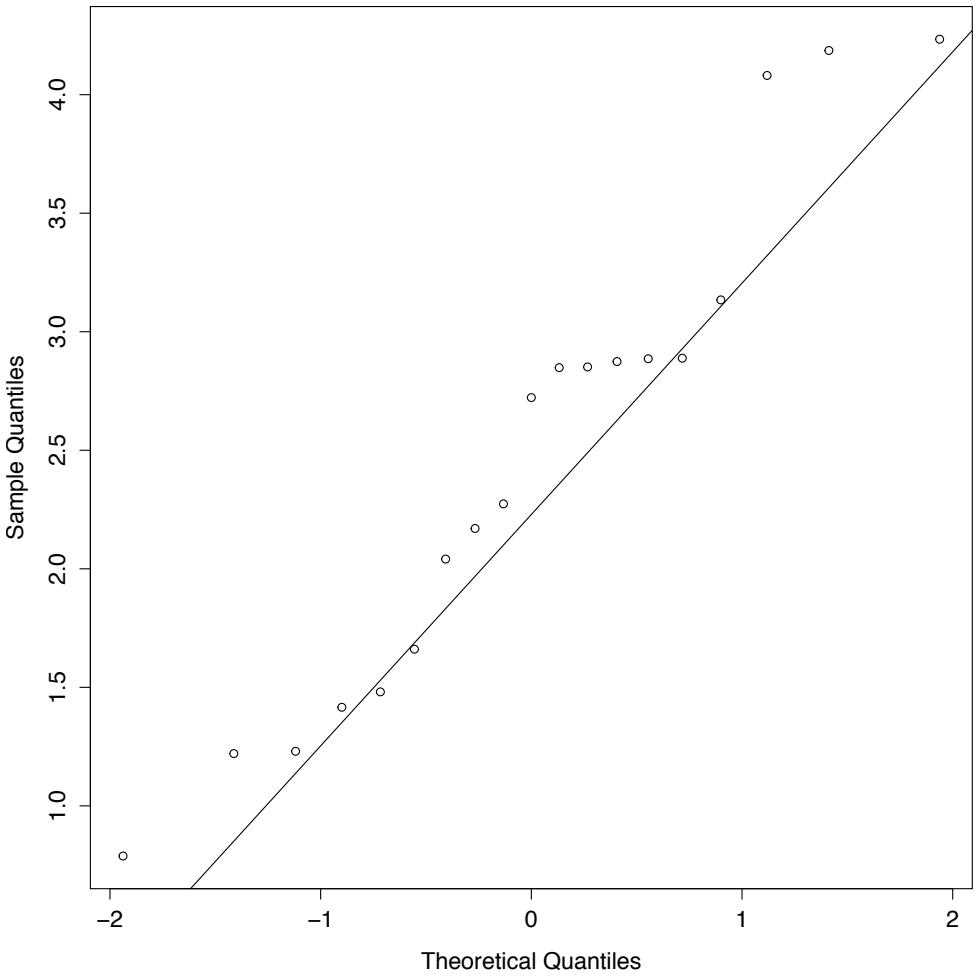

Figure S2. QQ-plot of HSF TF transcription levels under normal conditions. Normal distribution quantiles plotted against sample quantiles for the log-transformed data collected from non-stress shown as black circles. Black diagonal indicates perfect concordance.

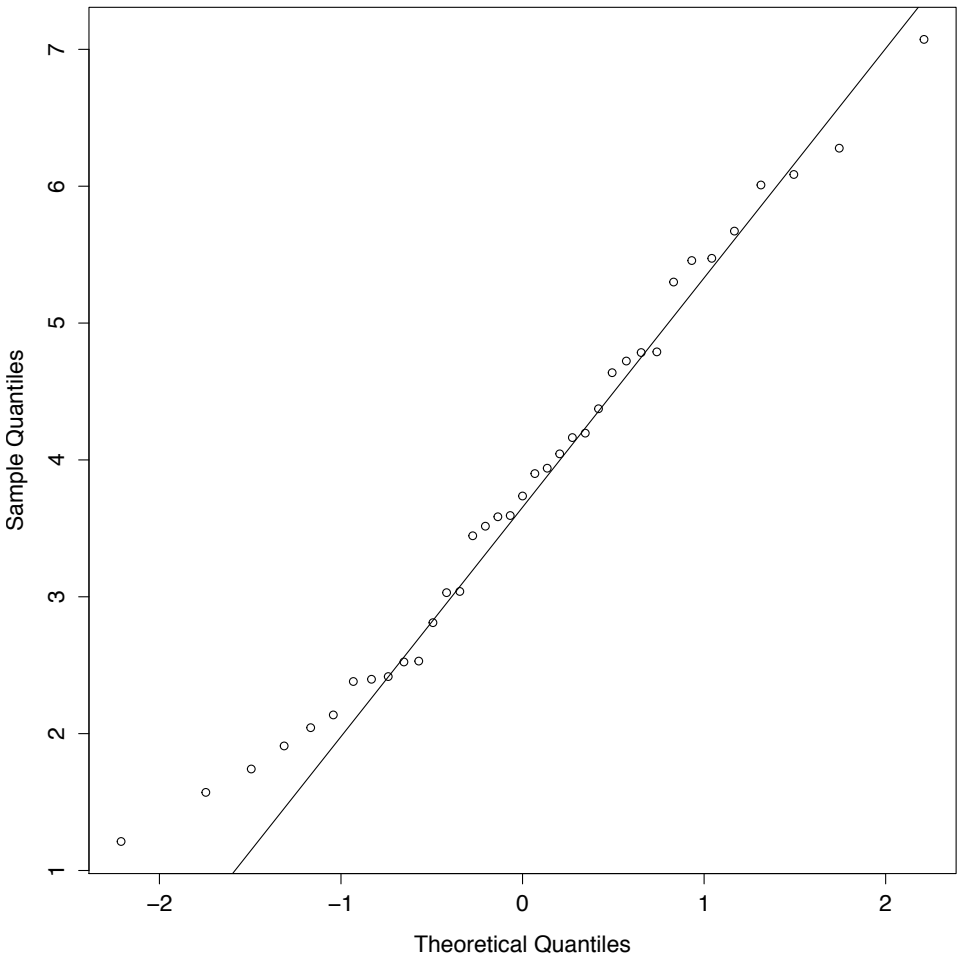

Figure S3. QQ-plot of HSP binding gene transcription levels under normal conditions. Normal distribution quantiles plotted against sample quantiles for the log-transformed data collected from non-stress shown as black circles. Black diagonal indicates perfect concordance.

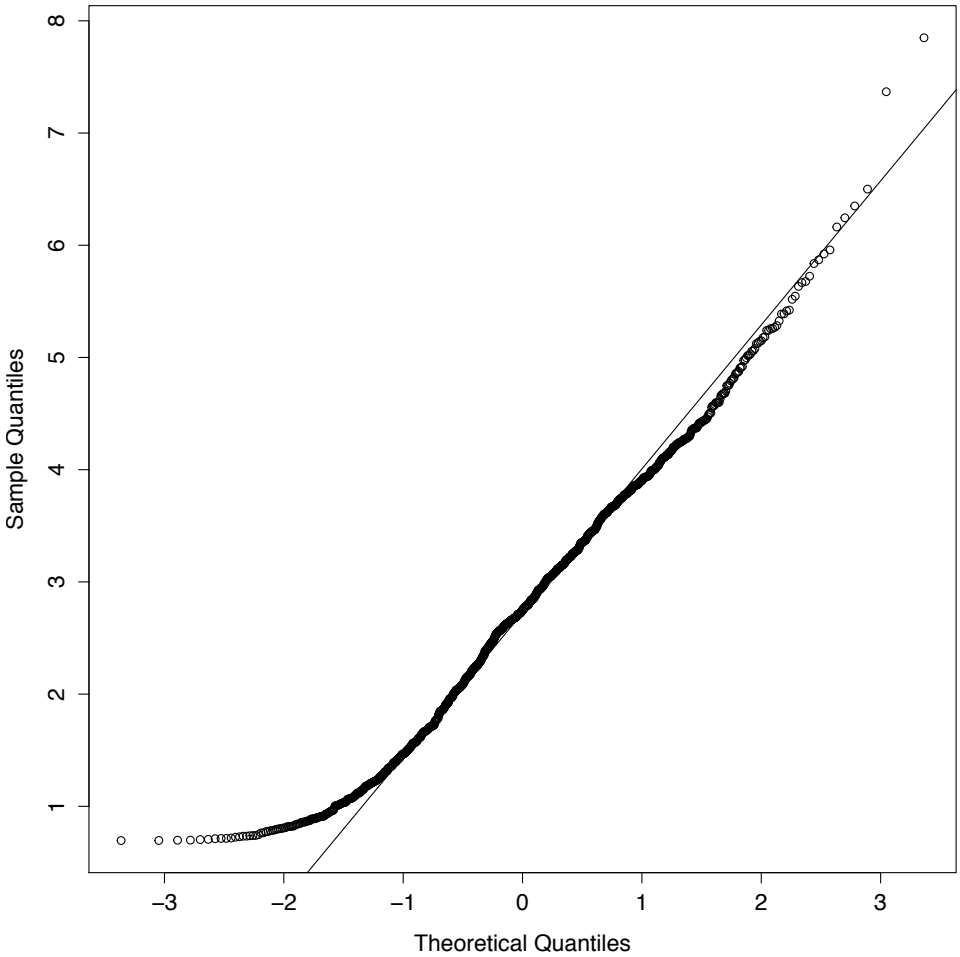

Figure S4. QQ-plot of other TF gene transcription levels under normal conditions. Normal distribution quantiles plotted against sample quantiles for the log-transformed data collected from non-stress shown as black circles. Black diagonal indicates perfect concordance.

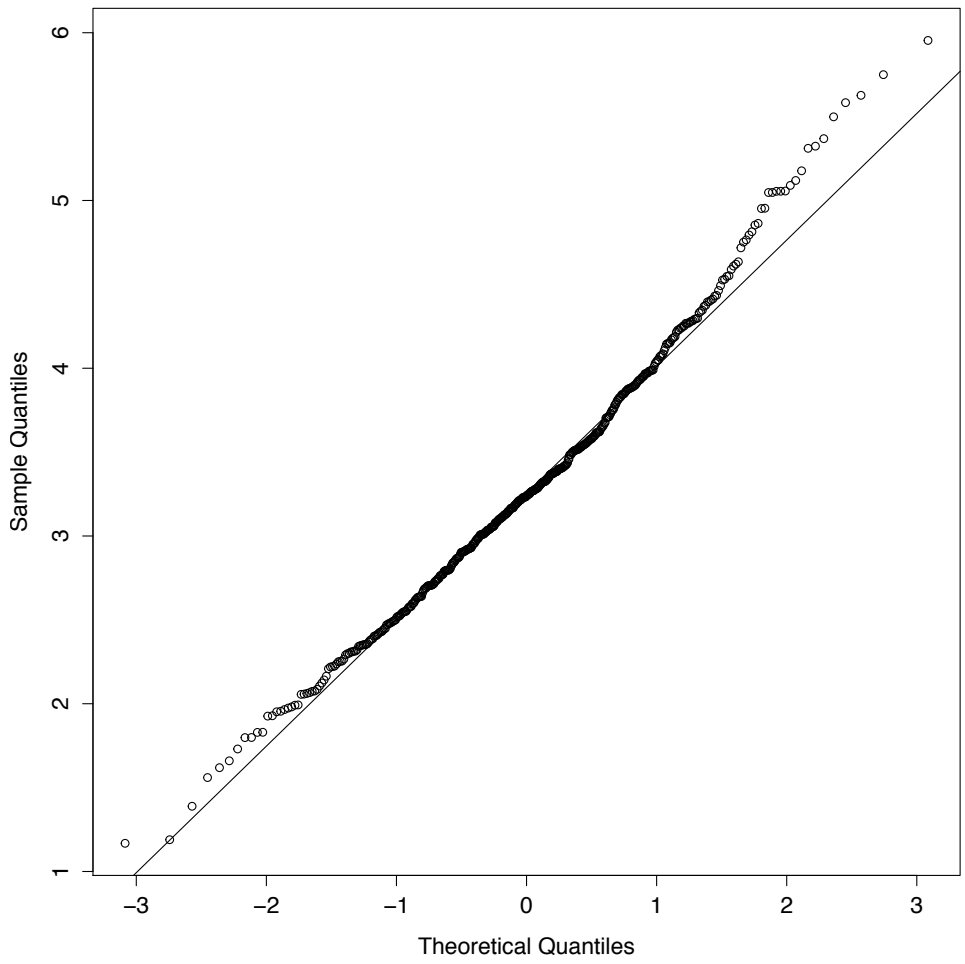

Figure S5. QQ-plot of housekeeping transcription levels under heat stress conditions. Normal distribution quantiles plotted against sample quantiles for the log-transformed data collected from heat stress shown as black circles. Black diagonal indicates perfect concordance.

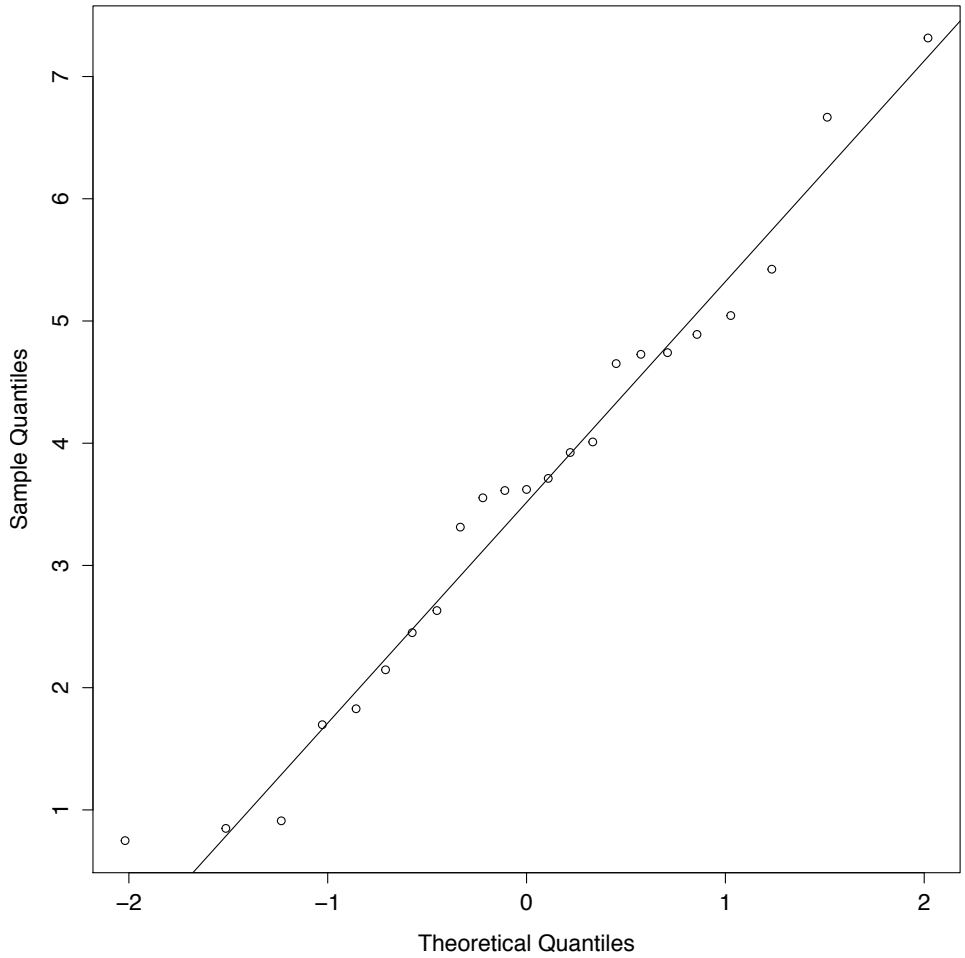

Figure S6. QQ-plot of HSF TF transcription levels under heat stress conditions. Normal distribution quantiles plotted against sample quantiles for the log-transformed data collected from heat stress shown as black circles. Black diagonal indicates perfect concordance.

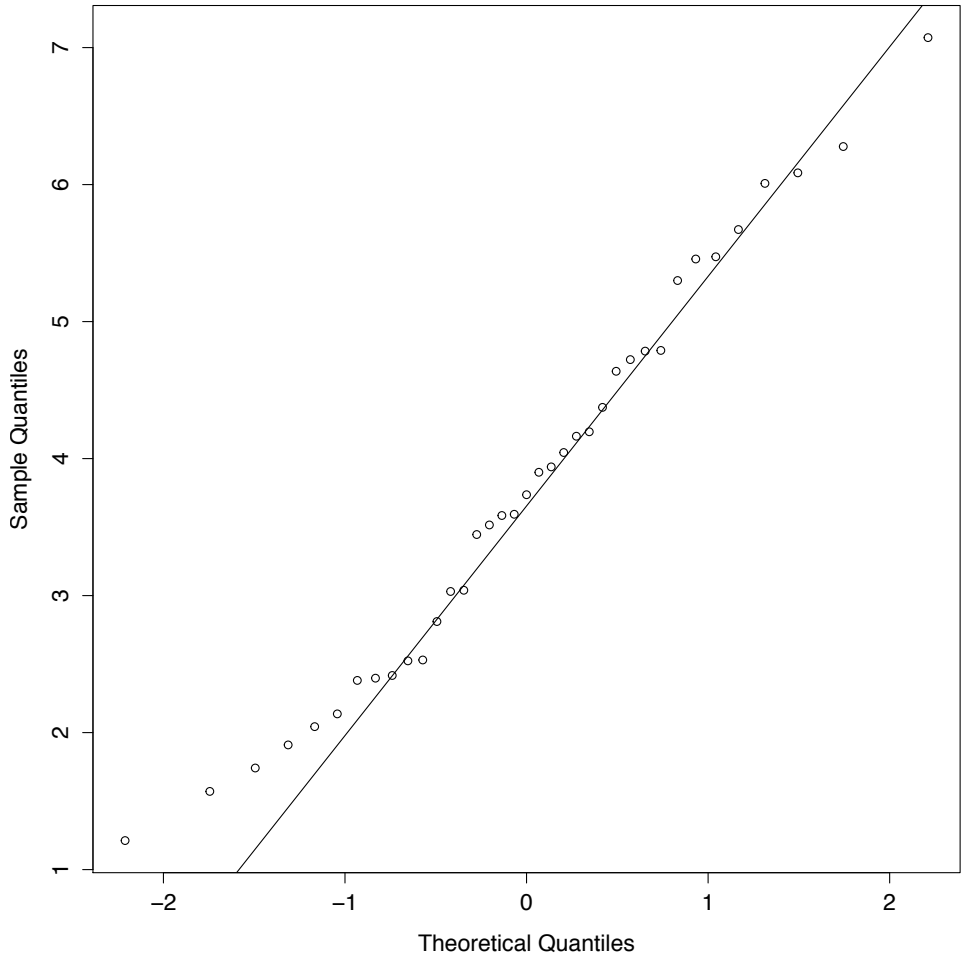

Figure S7. QQ-plot of HSP binding gene transcription levels under heat stress conditions. Normal distribution quantiles plotted against sample quantiles for the log-transformed data collected from heat stress shown as black circles. Black diagonal indicates perfect concordance.

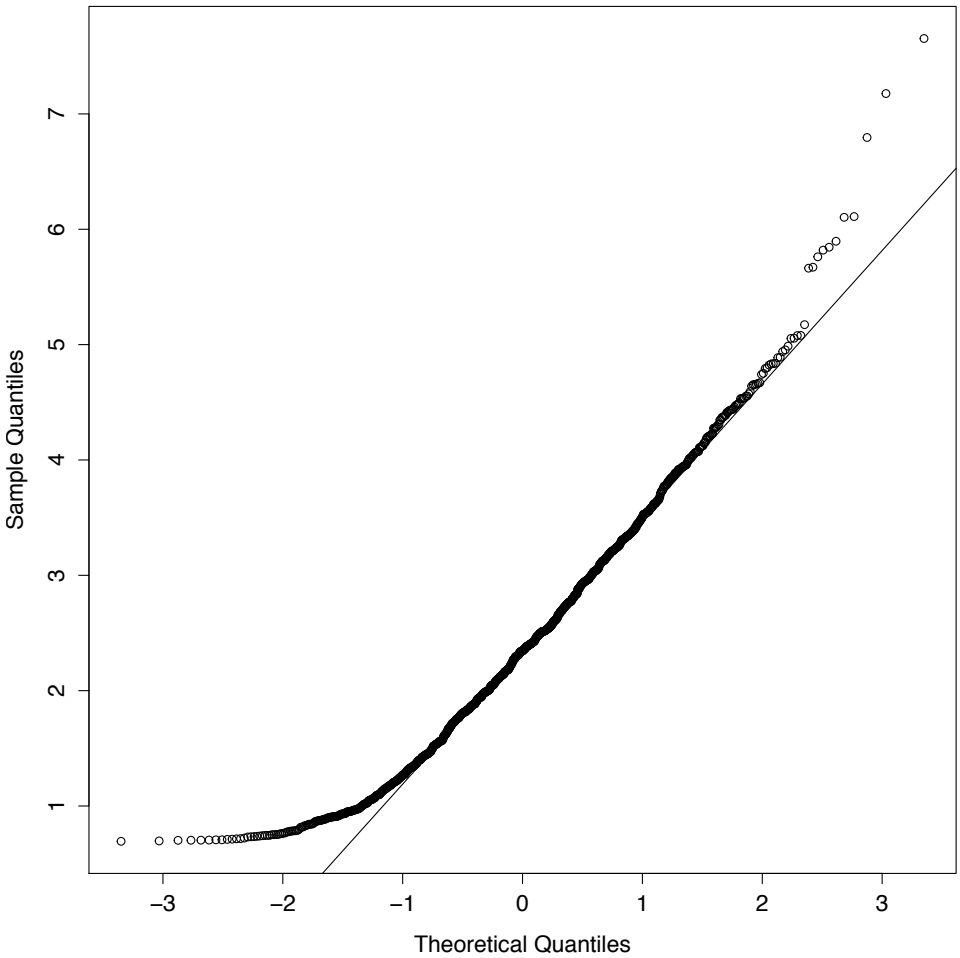

Figure S8. QQ-plot of other TF transcription levels under heat stress conditions. Normal distribution quantiles plotted against sample quantiles for the log-transformed data collected from heat stress shown as black circles. Black diagonal indicates perfect concordance.

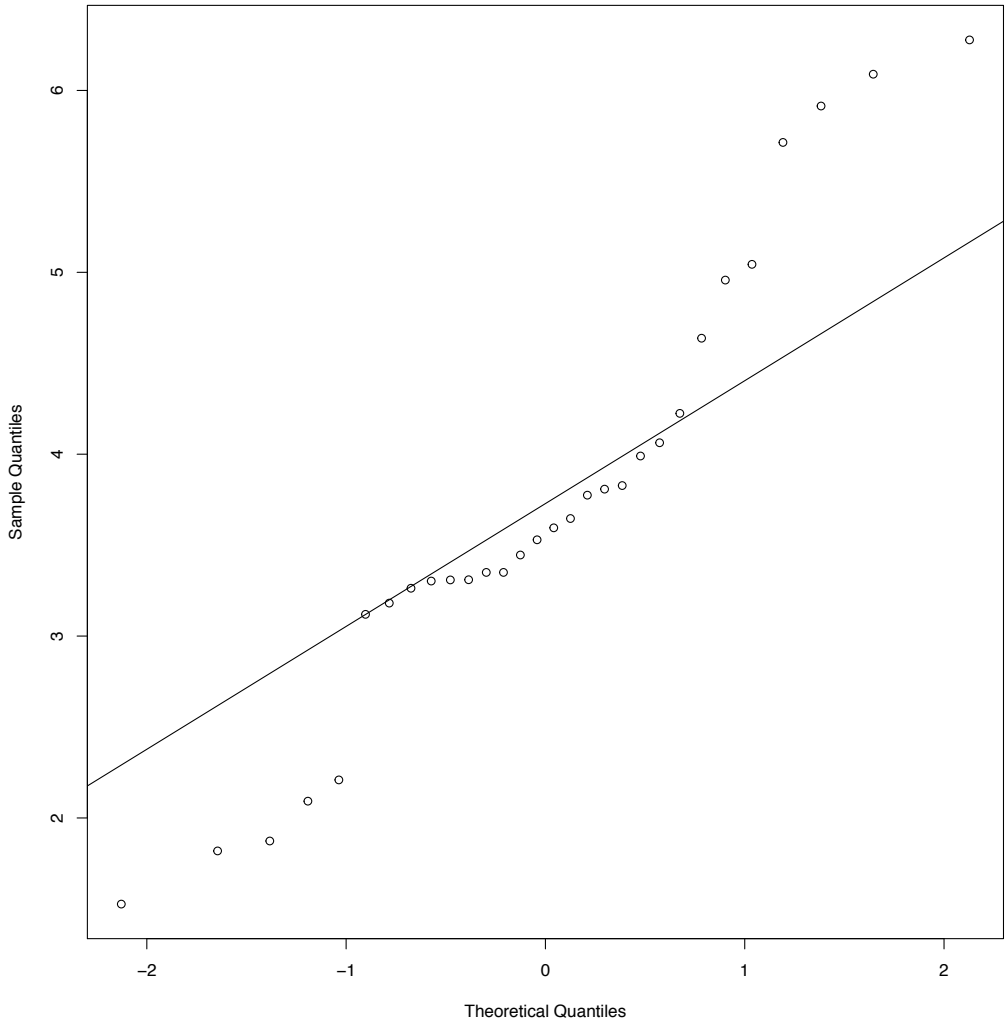

Figure S9. QQ-plot of HSP gene expression levels under normal conditions. Normal distribution quantiles plotted against sample quantiles for the log-transformed data collected from heat stress shown as black circles. Black diagonal indicates perfect concordance.

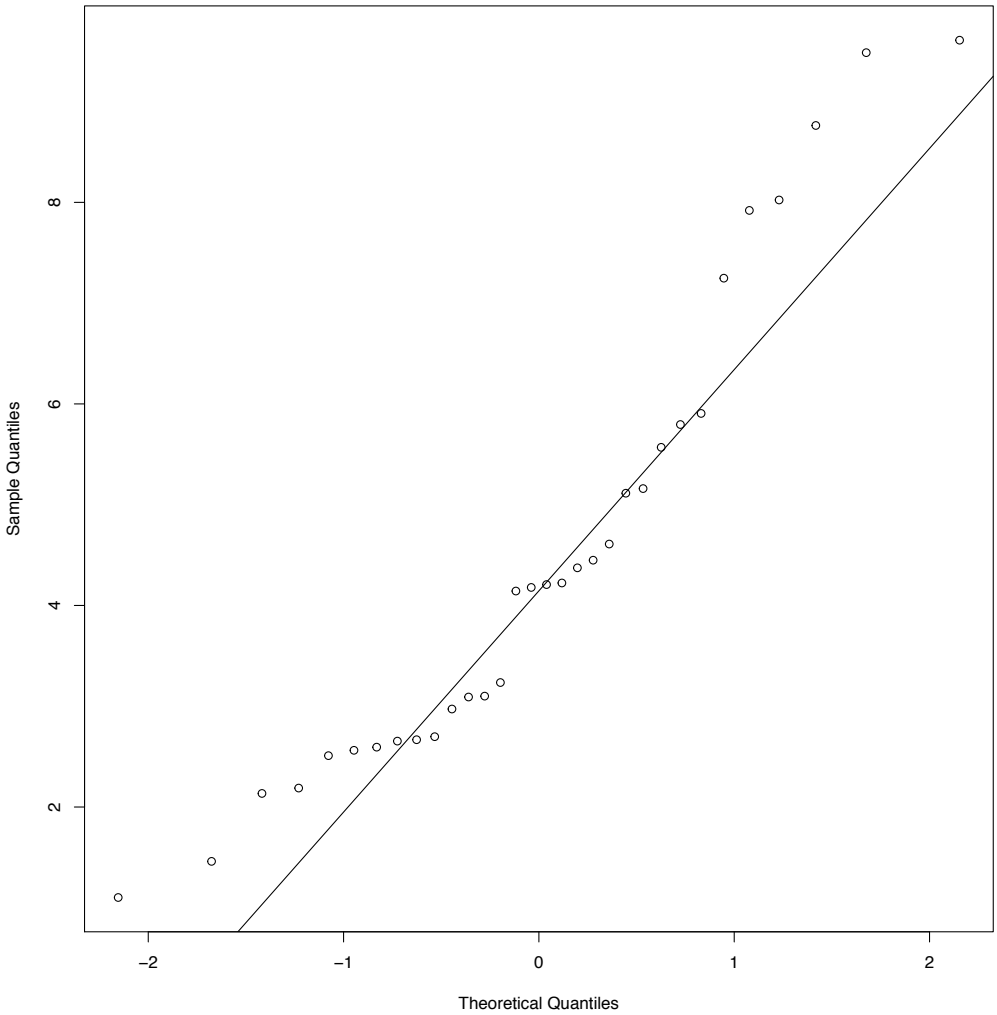

Figure S10. QQ-plot of HSP gene expression levels under heat stress conditions. Normal distribution quantiles plotted against sample quantiles for the log-transformed data collected from heat stress shown as black circles. Black diagonal indicates perfect concordance.