A Bioinformatics Pipeline to Explore Transcriptional Regulation in Plants

The activation of transcription via signal transduction pathways is one of the most sophisticated molecular mechanisms plants have to cope with and adapt to stressful environments. Scores of signal transduction pathways can be initiated depending on the combination of stresses experienced by the plant. Directly or indirectly, plant transcription factors (TFs) sense and respond to external signals such as light and temperature, as well as endogenous signals such as hormones. Moreover, TFs regulate other TFs, resulting in complex regulatory networks controlling thousands of genes in response to the various environmental signals. It is difficult to conceptualize genome-wide transcriptional regulation and even more challenging to organize, analyze, and visualize data at this scale. Network analysis of gene expression data is a popular way for plant scientists to deal with “-omic” scale data. However, the tools and techniques needed for such an analysis are not commonly taught alongside other plant biology curricula. Here, we present a flexible learning module that provides students with training in the construction and analysis of a coexpression network in the context of transcriptional regulation by TFs. This module can be taught as a traditional lecture or as a hands-on module for small lecture or laboratory courses. The first part of the lesson provides background and theory for the analysis, and the second part provides two step-by-step tutorials for hands-on exploration of the tools used for transcriptional analysis.

PART 1. BACKGROUND AND THEORY

Introduction to Gene Regulation, Transcription, and Transcription Factors

Gene expression dictates what functions a cell will have and allows organisms to respond to environmental signals. Plants and animals experience gene expression every day, and we can even see it happening. For example, UV light induces the expression of pigmentation genes in humans. Gene expression can be regulated at every step along the way from DNA to protein. The first step in gene expression is transcription, in which a mRNA molecule (mRNA) is synthesized from a DNA template by RNA polymerase. This lesson focuses on transcriptional regulation, but there are several additional mechanisms that regulate gene expression posttranscriptionally, translationally, and posttranslationally.

This lesson and the hands-on activities focus on the contributions of transcription factors to gene expression. Transcription factors are proteins that bind to a gene’s promoter region, or a more distant enhancer region, and dictate if that gene will be turned “on” or “off.” Within promoters, TFs recognize a specific regulatory sequence called a cis-regulatory element (CRE), also known as a regulatory motif. The binding of a TF protein to a CRE can either activate or prevent transcription from occurring, by either recruiting or blocking RNA polymerase, respectively. Gene promoter regions can have multiple copies of the same CRE or a variety of CREs. The interaction between TF proteins and their target sequences are quite specific; TFs can have up to 10^6-fold higher affinity for their target CRE sequences than for other DNA sequences.

Methods to Identify Transcription Factor Binding Sites

DNA sequences that serve as transcription factor binding sites can be identified using a number of experimental techniques designed to study protein-DNA interactions.

Historically, in vitro assays such as electrophoretic mobility shift assay and DNA footprinting have been used. In electrophoretic mobility shift assay (also known as “gel shift” assay), labeled DNA is mixed with proteins and then electrophoresed through a gel. If the DNA is bound by a protein, the mobility decreases and the band position is shifted. DNA footprinting is another assay used to explore protein-DNA interactions and identify the DNA sequence bound by the protein. When done in vitro, a purified TF protein is incubated with a radiolabeled DNA probe and then this mixture is treated with an endonuclease that cleaves phosphodiester linkages in DNA, resulting in polynucleotide fragments. Where a TF is bound to the DNA, it protects the DNA from nuclease cleavage. Cleavage patterns with and without TFs are compared by gel electrophoresis to reveal the TF binding site sequence. This approach has been expanded to the genomic scale and is called digital genomic footprinting (DGF), which is also known as DNase-seq. DGF can be performed in vivo to identify genome-wide binding sites for a TF of interest. In Arabidopsis thaliana, DGF was used to map 700,000 sites occupied by transcription factors in response to heat shock and photomorphogenesis, which revealed important TF regulatory circuits triggering these specific responses.

With the advent of high-throughput sequencing, technologies have been invented that allow analysis of TF-DNA binding sites and motifs in vivo and/or genome-wide. These technologies include chromatin immunoprecipitation followed by DNA sequencing (ChIP-seq) and, more recently, DNA affinity purification sequencing (DAP-seq). ChIP-seq enables the selective enrichment of DNA sequences bound by a TF protein of interest. TF-DNA complexes are chemically cross-linked prior to shearing of the DNA into small fragments and immunoprecipitation of the protein via antibodies attached to beads. Next, the protein is unlinked from the DNA. The
DNA is then purified, sequenced, and mapped to a reference genome. ChiP-seq has been a powerful and high-throughput method of identifying TF binding sites. However, ChiP-seq is not easily scaled because it is dependent on the availability and effectiveness of antibodies. DAP-seq is an assay that uses an in vitro expressed tagged TF to interrogate naked genomic DNA fragments to find binding locations and sequence motifs. The TF protein is bound to ligand-coupled beads and then allowed to interact with genomic DNA fragments. Unbound DNA gets washed away, while the bound fraction is eluted, amplified, and sequenced. Mapped reads are used to identify TF binding sites and motifs. This high-throughput assay was recently performed in Arabidopsis and resulted in the identification of 2.7 million binding sites for 529 TFs. Finally, as more information is learned about transcription factors and their binding targets and effects, computational methods can be used to predict additional interactions, which can then be experimentally verified.

Altogether, these technologies have created an atlas of thousands of TF binding sites and motifs. Comprehensive lists of TFs and their cis-regulatory elements are stored in a number of databases that have become invaluable resources for biologists studying transcriptional regulation. The most complete databases include PlantTFDB 4.0, AGRIS (Arabidopsis Gene Regulatory Information Server) and RARTF (RIKEN Arabidopsis Transcription Factor database). These databases catalog experimentally determined as well as computationally predicted CREs and organize TFs into families. Knowledge about these plant TFs and the sequences they bind to are used in identifying gene expression networks, as described below. Additionally, experimentally determined interactions can be extrapolated computationally to infer other interactions, which can then be experimentally verified.

Conserved Families of Plant Transcription Factors

Transcription factor proteins are organized into families based on several criteria including the presence, absence, and linear organization of multiple different protein domains, their phylogenetic relationship to known proteins, and the sequence motif of their DNA binding domains. Some TF families are characterized by their involvement in plant response to stress, such as extreme temperature, high or low light, water deficiency, high salinity, and pathogen infection. For example, the basic leucine zipper (bZIP) family is one of the most diverse TF families. All bZIPs have a conserved bZIP domain, which is 60 to 80 amino acids long and comprises a basic region and a leucine zipper region that are structurally and functionally different. bZIP TFs regulate a number of processes including response to drought, cold, sugar signaling, nitrogen signaling, and light response.

Plants have more transcription factors and a wider diversity of transcription factor families than other organisms. Drosophila melanogaster and Caenorhabditis elegans have ~600 genes that encode transcription factors, which make up 4.7% and 3.6% of their genomes, respectively. The model plant Arabidopsis has 1318 loci widely recognized as encoding TFs; however, some research has suggested that Arabidopsis has up to 2620 TF loci, making up 10% of its genome. These TF genes are organized into 72 families. About half of Arabidopsis TFs have plant-specific DNA binding domains, including WRKY, AP2/ERF, and NAC transcription factors, while the other half include families that are conserved across eukaryotes, such as bZIP, Myb, and bHLH. Furthermore, some plant-specific TF families are also present in green algae, and some are restricted to vascular plants.

Transcription Factors are Key regulators for Diverse Functions: Five Case Studies

PHR1

The transcription factor PHR1 regulates a large number of genes in response to phosphate (Pi) deficiency. PHR1 binds to P1BS, the PHR1 binding sequence, which is found in the promoter regions of many genes that are induced by Pi starvation. Plants in which PHR1 expression is downregulated have an impaired response to Pi starvation. When the P1BS motif of Pi-responsive genes is mutated, the genes are no longer expressed during Pi starvation. In addition to its key role in Pi-deficiency responses, this transcription factor is also involved in adaptation to high-light conditions when phosphate is limiting.

NACs

The role of the NAC family of plant-specific transcription factors in the adaptation of plants to land has been studied using both loss-of-function and gain-of-function studies. Xylem vessel elements provide structural support to plants and facilitate efficient water conduction. In the moss Physcomitrella patens, the triple NAC gene deletion mutant ppvns1/6/7 has a water transport deficiency due to abnormalities in structural elements called hydroids and stereids. Overexpression of PpVNS genes in either P. patens or Arabidopsis results in cell death. Collectively, these studies reveal that these NAC TFs regulate hydroid cell differentiation by inducing their death and regulate stereid cell differentiation by promoting cell wall thickening. Studies of NACs showed that the evolution of programmed cell death facilitated water conduction in plants, which contributed to their adaptation to land.

LBDs

The LATERAL ORGAN BOUNDARY DOMAIN (LBD) gene family of transcription factors were first identified as developmental regulators expressed at the boundary of lateral organs. A microarray analysis of genes transcribed in response to nitrogen deprivation or resupply identified the genes encoding LBDs as nitrogen-responsive, suggesting a role in nitrogen metabolism. Additional studies showed that many of the genes significantly upregulated in plants overexpressing LBDs are also responsive to nitrogen in wild-type plants, corroborating the role for LBDs in the nitrogen response.

WRKYs

They WRKY transcription factors (named for a conserved set of amino acids and pronounced “worky”) are a very large family in plants with many different functions. In the second part of this lesson, students can examine gene regulation from a microarray study of wild-type and wrky1 mutant plants in response to nitrogen treatment.
RBE

RABBIT EARS (RBE) is a transcription factor that was identified through the phenotype of a loss-of-function mutant, in which the flower petals form abnormally. In the second part of this lesson, students can examine gene regulation from an RNA-seq study of wild-type and plants transiently overexpressing RBE.

Functional Testing of the Regulatory Role of Transcription Factors

The regulatory role of a transcription factor can be found using functional approaches. The most common approaches used to study TF function are loss-of-function and ectopic expression studies, in which the expression of the gene encoding the TF under study is either partially or wholly silenced or overexpressed and the phenotype of the affected plants is analyzed. These methods are combined with transcriptomic studies to identify differentially expressed genes (DEGs) that are direct or indirect targets of the TF under study. The second part of this lesson provides step-by-step guides so that students can carry out such analyses.

There are many ways to generate loss-of-function mutants by deleting, interrupting, or mutating part of the coding or promoter sequence of a gene of interest. Some techniques include T-DNA insertion, virus-induced gene silencing, RNAi, and point mutations.

Overexpression experiments can also result in abnormal phenotypes that reveal insights about the function of a wild-type gene. Often, overexpression studies are used in parallel with loss-of-function studies. The most common method used to overexpress a gene is transformation of plants with Agrobacterium tumefaciens containing a plasmid in which the gene of interest is under the control of a strong, constitutive promoter, such as the CaMV 3SS promoter. It is also possible to switch on TFs selectively by choosing promoters that promote expression only in a particular cell type or at a specific developmental stage.

Additional insights about transcription factor function can come from studies to determine where and under what conditions the transcription factor itself is expressed. As an example, a family of transcription factors that are transiently upregulated by cytokinin were identified and named CYTOKININ RESPONSE FACTORS (CRFs). Loss-of-function analysis revealed phenotypes consistent with cytokinin insensitivity. Transcriptomic analysis showed that many cytokinin-regulated genes are differentially transcribed in the crf mutants, demonstrating that they contribute to the transcriptional response to cytokinin.

Differential Expression Analysis: Tools, Techniques, and Outcomes

This section provides a general overview of the steps involved in carrying out an analysis of differential gene expression. Subsequent sections will describe how the DEGs are functionally characterized and guide student through step-by-step, hands-on exercises starting with data sets obtained from microarray analysis and RNA-seq.

Mutant and overexpressor lines can be used to determine the genome-wide regulatory role of a transcription factor protein on target gene expression using transcriptomic technologies. The most common technologies are DNA microarrays and high-throughput RNA-seq. Both of these methods measure the level of mRNA in each sample, and both start by converting mRNA into cDNA through the process of reverse transcription. Statistical analyses are then performed on the expression data to determine significantly differentially expressed genes. In the context of a TF mutant, the DEGs are likely direct or indirect targets of the TF. The biological context of the results can be examined using a number of bioinformatics tools. In this section, the general process of gene expression analysis is explored.

Data Collection: Microarrays and RNA-Seq

There are two standard methods for obtaining transcriptomic data for analysis: microarrays and RNA-seq. In this section, we will describe each method and how the data obtained from each method is preprocessed and analyzed.

cDNA Microarrays

Microarrays are glass slides that contain thousands of short DNA probes corresponding to known sequences in an organism’s genome. Sample cDNA labeled with a fluorescent dye is hybridized to the microarray chip. The amount of hybridization (as determined by level of fluorescence) for a particular probe is proportional to the number of cDNA fragments in the sample.

Two types of microarrays are available for expression analysis: two-color and one-color (or single-channel) microarrays. Two-color arrays facilitate comparison of two samples on the same chip. Here, the two different samples are labeled with two fluorescent dyes with different fluorescence emission wavelengths. The two labeled cDNA samples are mixed prior to hybridization to the chip. The relative intensities of each dye can be analyzed to determine the ratio of genes that are up or downregulated in the samples. Microarrays are not able to provide absolute values of the number of transcripts per cell without a calibrated reference sample. One-color arrays estimate the relative level of gene expression within a single sample. For comparison studies, every sample gets its own chip and relative expression values are compared. Most researchers prefer single-channel arrays due to more robust results and more straightforward data analysis.

RNA-Seq

RNA-seq technology takes advantage of next-generation sequencing of cDNA molecules. This technology measures the number of sequenced fragments that map to a particular transcript, which is proportional to an mRNA’s abundance within the sample. RNA-seq has several advantages over hybridization approaches such as microarrays. For example, RNA-seq has higher sensitivity compared with microarray technology and thus is able to detect differences among very low and very highly expressed genes. Absolute quantification of highly expressed genes is limited in microarray technology because the upper limit for fluorescence is determined by number of probes on the chip.
Likewise, some lowly expressed genes do not appear in microarray outputs because the level of fluorescence does not pass the minimum threshold. Another advantage of RNA-seq is that its outputs are at the level of single-nucleotide resolution, which can provide information about alternative splicing that would not be possible with microarrays.

A typical RNA-seq experiment involves preparation of the sample library and experimental design that aims to generate sufficient sequencing depth to capture changes in gene expression. Sequencing may be done using single-end or paired-end reads, where sequencing is performed at one or both ends of the cDNA fragments, respectively. Given the same read length, paired-end reads provide two anchoring points and hence often result in more precise alignment of read sequences to a reference genome. This improvement in alignment precision is particularly useful in genomes with multiple near-identical copies of genes (e.g., maize and soybean).

Despite all of the advantages of RNA-sequencing over hybridization technologies, a major challenge is the difficulty in processing and analyzing outputs. Traditionally, the alignment process required extensive computing power and could not be performed on an average laptop or desktop computer. However, recent advances in mapping algorithms coupled with ever-improving computing power have made this possible as well. Likewise, many software tools have been developed to analyze RNA-sequencing data, and the researcher must choose those that are most appropriate for their study.

Analysis of Transcriptomic Data

The results of RNA quantification technologies can be used to understand transcriptional regulation by comparing the transcriptomes of wild-type plants and those overexpressing or underexpressing a specific transcription factor. DEGs can be profiled using a series of computational steps, including (1) quality control, (2) data preprocessing and normalization, and (3) differential expression analysis followed by in silico clustering and functional analysis to infer hypotheses about biological significance. These general steps are similar for microarrays and RNA-seq; however, the algorithms and tools are different for the specific technology.

Quality Control

Quality control, or QC analysis, is designed to identify technical errors that occur during sample preparation, processing, and data analysis and should be performed at each of these steps. Poor sample and/or data quality can arise from variability in a number of experimental steps, including but not limited to experimental design, efficiency of nucleic acid extraction, and/or processing for downstream analysis, image processing of microarrays, and QC bias in raw RNA-seq reads. By following QC workflows, researchers can get an idea of where things went wrong and when to remove bad samples. Quality control measures performed early in the pipeline can determine if extracted RNA is of adequate concentration and integrity prior to downstream processing or if those samples should be re-done or discarded.

QC software has been developed for both microarray and RNA-seq technologies to assess hybridization or sequence quality, respectively. Most QC software output results as easy-to-follow visuals that give a bird’s-eye view of sample distribution and features to help researchers identify potential sources of noise in the data. In RNA-seq analysis, quality assurance is imperative when evaluating acquired raw read counts and aligning reads. These steps are done prior to normalization and help the researcher determine if unacceptable errors occurred during sample preparation or during the sequencing run.

A common source of poor data quality is a failed array hybridization or a failed RNA-seq library. In these scenarios, the affected array or library will often have aberrant global measure of gene expression and can therefore be identified using these global measures. A failed array hybridization would have low intensities across all probes, while a failed library preparation would result in a low sequencing yield (i.e., fewer sequencing reads). An early QC step to capture such aberrant data is to apply a clustering method to the raw intensity/read count data. In such clustering approaches, the samples derived from the same biological condition should be more alike than those from other conditions, and low-quality data sets whose global gene expression profile does not match any of the other samples within the same experiment are well separated from all the other samples. Any such low-quality data sets identified through this analysis should be excluded from all further analysis. In RNA-seq analysis, the QC step of identifying and eliminating failed libraries is typically done after the preprocessing steps because the RNA-seq reads need to be converted into gene expression counts to allow meaningful comparison of samples.

Data Preprocessing and Normalization: cDNA Microarrays

Data preprocessing attempts to remove technical differences to observe biological differences. Common preprocessing steps include data normalization, localization, standardizing, and mean-centering. Data preprocessing is used to transform raw data from transcriptome analysis to allow comparisons among samples. Preprocessing of data from cDNA microarrays is often needed to account for technical biases introduced across a series of hybridizations. For example, normalization algorithms are used to correct for variances in the fluorescent intensity of signals in microarrays.

Data preprocessing is usually the first step in gene expression analysis; however, it is important to evaluate if the chosen preprocessing method is appropriate for the transcriptome experiment of interest. Some data preprocessing, such as normalization, can mask significant biological differences or contribute to false positives, and different methods can impact downstream differential expression analysis because they transform the data in different ways, which impacts how variation is measured within and between samples.

The general steps of data preprocessing include (1) background correction, (2) standardizing scale unit based on either log transformation or overall brightness, (3) adjustment characteristics (based on platform used), and (4) summarizing information from several spots into a single measure for each gene. Once these corrections for various technical issues have been made, microarray data from multiple arrays are normalized to render them comparable.
Several normalization methods are available for microarrays, such as quantile normalization. Rather than simply normalizing the average intensity value of each data set, the data sets are normalized so that the 25th and 75th highest intensity values are also normalized. This and other popular normalization methods rely heavily on generalized assumptions about expression data (e.g., most genes are not differently expressed, or all samples have similar distributions of gene expression) and can lead to incorrect analysis especially in data sets in which the assumptions do not hold up.

Data Preprocessing and Normalization: RNA-Seq

Data from RNA-seq also requires preprocessing to account for within and between sample variations caused by differences among library sizes, GC contents, or even mapping method. To compare gene expression between two conditions from an RNA-seq experiment, effects not associated with the biology must be eliminated. These effects are due to technical errors as well as sample library characteristics such as the varying number of sequenced reads that map to a gene of interest, which is dependent on sequence length (in other words, genes that are longer are more likely to have “hits” in the RNA-seq sample, simply due to their longer length). Other sources of variation specific to RNA-seq data sets are PCR duplicate reads, variable sequencing quality, residual adapter sequences, etc. Data preprocessing steps, especially trimming of the RNA-seq reads and quality filters, alleviate these sources of spurious variation prior to mapping reads to the genes. These mapping data are then converted into gene expression counts by summing up the total number of reads that align to any given gene.

In RNA-seq analysis the QC step of comparing global gene expression profiles is typically performed after the reads are aligned to the genome and the gene expression counts are summarized. At this stage, the clustering approaches described earlier can identify any RNA-seq libraries whose global gene expression profiles is aberrant i.e., is an outlier from all other samples or groups with the wrong set of replicates. Such aberrant libraries are considered to have failed at the experimental stage and should be excluded from further analysis.

Similar to the intensity values in microarrays, gene count data from RNA-seq are also normalized to render the counts comparable across samples. Common normalization methods include trimmed mean of M-values (TMM), the DESeq package, quantile normalization, and RPKM (reads per kilobase per million). Care must be taken when choosing a normalization method because the resulting normalized data will impact the results of subsequent differential expression analysis. The combination of normalization, filtering, and differential expression analysis strategies can have widely varying results on certain data sets, so researchers are encouraged to perform multiple comparisons to determine the robustness of each analysis pipeline. For example, a study using RNA-seq data from 726 individual Drosophila samples tested multiple analytical pipelines with various combinations of three different filtering strategies, eight normalization methods, and two statistical methods for differential expression analysis. This study found that the most robust pipeline involved (1) normalization using DESeq, (2) filtering by removing low-expressed genes after normalization and data distribution fitting procedures, and (3) statistical analysis by generalized linear model assuming a negative binomial distribution. In addition, they found that most normalization methods were robust to filtering but that the TMM and quantile normalization methods were particularly sensitive to removal of low-expressed genes.

Differential Expression Analysis: Statistical Methods

Many methods are available to identify differentially expressed genes between treatment and control experiments. As an example, when studying the effects of a perturbation (overexpression or knockdown) of a transcription factor protein, mutant, and wild-type transcriptomes are compared with identify genes that are mis-regulated in the mutant. Obviously any two transcriptomic data sets will have differences. Finding meaning in these differences requires the use of statistics.

Statistical tests determine the probability of the null hypothesis ($H_0$), which is that the differences between samples are due to random noise and that the two samples are representative of the same underlying distribution. The most robust statistical models provide a probability value (P value) or other metric (Z score, q-value, etc.) that the read counts observed among the samples are from the same distribution. A small P value suggests that the null hypothesis is not true, and a gene of interest is differentially expressed due to biological reasons and not due to random noise in the data.

An appropriate statistical test must be chosen for the experimental design and the platform used to determine gene expression. For one-variable experiments (e.g., wild type versus mutant or overexpressor), simple pairwise comparisons may be sufficient. Common statistical tests that compare two conditions include log ratio, f test, and rank product. For multivariate data and experiments with more complex structure, more robust tests are necessary and can include ANOVA for normally distributed data and generalized linear models for non-normal data. A commonly used statistical package that implements generalized linear models is limma (limma stands for linear analysis of microarray data, as it was originally developed for use on microarray data).

Considerations for Analysis of RNA-Seq Data

A unique feature of RNA-seq data compared with microarray data is that transcripts are reconstructed based on mapping of reads to a reference genome or transcriptome (unless de novo assembly is required). Expression is then based on the number of read counts that map to a particular gene. Gene expression values are expected to follow a Poisson distribution, where the variance is equal to the mean. However, due to the high dynamic range of RNA-seq data, it often suffers from overdispersion, i.e., the observed variance in read count is greater than the assumed variance from the Poisson distribution. When overdispersion occurs across samples, the gene counts are better estimated by a negative binomial (NB) distribution, which allows the variance to be bigger than the mean. This NB distribution has become the basis for many of the most popular RNA-seq analysis software packages including Cufflinks, EdgeR, and DESeq. For example, the DESeq package contains models to estimate the variance-mean
dependence in read-count data and test for differential expression using the NB distribution in generalized linear models.

**Multiple Comparison Testing**

After DEG analysis, additional statistical methods are needed when comparing more than one pair of means in a data set. A typical RNA-seq study reports the expression levels of tens of thousands of genes, and each gene is tested for differential expression at a predetermined level of significance (P value). When testing multiple hypotheses simultaneously, the probability of observing a significant result due to chance increases greatly (i.e., the false positive rate increases). The higher false positive rate that often occurs in large scale genomic data sets can be controlled using additional statistical tests. This post-hoc analysis is called multiple comparison testing. Two common approaches are the Bonferroni correction and the false discovery rate, and both assume independent comparisons. Bonferroni correction is a stringent method that can lead to a high false negative rate in such large sets. Instead, the Benjamini-Hochberg procedure is used to control the false discovery rate. This method is appropriate for a large number of multiple comparisons, such as those from DEG analysis of microarray and RNA-seq data. While both of these procedures correct for potential false positives, they come with the risk of potentially increasing the number of false negatives, which may impede the discovery of biologically meaningful results. Therefore, researchers must carefully weigh the cost of false positive versus false negative results to determine an acceptable false discovery rate.

**Functional Analysis of DEGs**

**Clustering**

To make sense of the often large sets of gene expression data, clustering algorithms are applied to group genes with similar expression patterns into natural subgroups. Clustering is an effective way to identify patterns in large-scale expression data. The underlying assumption is that genes that act together or sequentially in the same process would be coexpressed. For example, two genes encoding different proteins that function in a complex may cluster together by their expression patterns. Genes that cluster together may also be coregulated by common transcription factors. Therefore, clustering often precedes the analysis of Gene Ontology (described below).

Hierarchical agglomerative clustering is the classic “bottom-up” method, in which the hierarchical agglomerative clustering algorithm builds a dendrogram where groups are nested and organized as a cluster tree; it begins by placing each individual gene in its own cluster then merging clusters based on degree of similarity (also described as “distance”). Distances between new and old clusters are recalculated and the merging procedure is repeated until all genes are grouped into a single cluster. The final dendrogram is a summary of the clustering procedure and provides a bird’s-eye view of the similarity among expression values in a DEG data set.

Quality threshold clustering (QTC) has no implied hierarchy of clusters and no defined measure for how distinct any two resulting clusters are. Instead, QTC optimizes the identification of the largest possible clusters that are within a user-defined distance. Resulting clusters include genes with similar expression patterns across samples.

Nearly all clustering algorithms rely on some distance metric, such as Euclidean distance (normal or squared), Spearman correlation coefficient, Pearson correlation coefficient, and Manhattan distance. This distance is a measure of the similarity in the expression profiles of any two genes, across all samples, such that more similar expression profiles are considered closer while dissimilar expression profiles are considered distant. Each metric computes distance in a different way; therefore, clustering outcomes can vary widely. For example, the Manhattan metric calculates the distance that would be traveled between two data points if following a grid-like path, while the Euclidean metric measures the “as-the-crow-flies” distance. Manhattan distance is the sum of the differences of the components between two data points, whereas Euclidean distance is the square root of the sum of the squares of the differences between corresponding values of two data points. Euclidean distance is useful for dense or continuous data, whereas Manhattan distance may be preferable for high dimensional data. Because of these differences, the researcher must decide what algorithm is most appropriate for their data and then carefully examine the results to see if they make biological sense.

**Gene Ontology**

A convenient way to assign biological function to a gene or group of genes is to use Gene Ontology (GO). GO is a set of structured, controlled vocabularies representing properties of gene products that provide information about their molecular function (e.g., binding activity), biological process (e.g., signal transduction), and cellular component (e.g., rough endoplasmic reticulum) domains. The Gene Ontology Consortium seeks to collect and categorize information about gene products for many species and makes these descriptions publicly available in databases for ease of use by researchers. GO terms are used to annotate gene products from different species, based on supporting experimental and/or computational evidence. As new evidence becomes available, GO is updated by biologist-curators.

GO is most commonly used to perform enrichment analysis on sets of genes that have been clustered in some way. Enrichment analysis uncovers GO terms that are over- or underrepresented within a given gene set, which is determined by the frequency of the GO term in a sample set compared with its frequency in a defined background set of genes (usually, but not necessarily, the entire genome). A P value is assigned indicating the probability that x number of genes out of the gene set are associated with a particular GO term, given the proportion of genes in the whole genome associated with that same term. The closer the P value is to zero, the more enrichment of genes with that GO term is for the given gene set.

It is important to note that one gene may be associated with several GO terms. The most common tests for overrepresentation analysis of GO terms are the Fisher’s exact test and the hypergeometric distribution, both followed by correction for multiple testing, like FDR (false discovery rate). Fisher’s exact test takes
values within a contingency table and determines the probability (represented as a P value) that the observed values occur under the null hypothesis that there is no association between categorical values. The Fisher’s exact test is best for small gene sets. The hypergeometric test analyzes each GO term independently of all other GO terms.

While GO is a convenient way to explore the biological interpretation of differential expression analysis, it is limited by the amount of available information for a particular species, meaning that not every gene in the genome is annotated or has a known function. As with any computational or bioinformatics analysis, it is important to think about biological experiments that can be conducted in the laboratory to gain further insight on the function of a gene(s) of interest.

**Network Analysis**

Another method used to explore the biological significance of a set of DEGs is network analysis. Network construction is a useful method to organize and analyze large -omic scale data sets with hundreds to thousands of data points, such as the results from a transcriptome study. In particular, networks are useful in understanding the regulatory role of a transcription factor on a whole system through construction of coexpression and/or gene regulatory networks (also known as GRNs).

Network analysis is based on graph theory, where a graph is defined as a pair \((V, E)\) where \(V\) is a set of vertices or nodes, and \(E\) is the set of edges connecting the nodes. Such graphs can be directed or undirected. For example, a graph might have a pair of nodes connected directionally (e.g., A goes to B) to represent a stepwise biological or metabolic process.

In a gene coexpression network, the graph consists of gene IDs represented by nodes and the correlated relationship between genes represented by undirected edges. The connections between gene nodes provide functional information about the relationship between those genes. It is widely accepted that interacting genes are more likely to share a similar function or be involved in a similar biological pathway or process, a principle known as guilt-by-association. Coexpression networks are generated by correlating the expression patterns of pairs of genes across biological conditions, in which two genes may have strong positive or negative (anticorrelation) correlation. It should be noted that coexpression and clustering for networks (e.g., WGCNA, weighted gene coexpression network analysis) assume the data are normally distributed. Therefore, appropriate data scaling and/or normalization methods should be applied before attempting network construction.

A gene regulatory network adds an additional layer of information to a coexpression network, in which protein-DNA interactions are included as directed edges, where a gene node that encodes a protein (source node) has a directional edge connected to another gene node (target node). GRNs are useful when analyzing data from experiments designed to test transcription factor function. These networks can provide insight about functional redundancy of regulatory factors when there are edges connecting transcription factor nodes to common target genes.

After a network has been constructed it can be analyzed to explore descriptive network properties such as degree (number of edges connected to a node), neighborhood connectivity (connectivity of neighbors), clustering coefficient (how nodes are connected in their neighborhood), and betweenness centrality (how much this node controls other nodes). The central premise of these analyses is that the position of the node within the network topology is related to the importance and/or influence of the node in the network. Broadly, the more central or connected a node is to the network the more influential it is expected to be. For example, a highly connected transcription factor node may indicate that it is a master transcriptional regulator.

Degree is the number of connections (edges) a node has to other nodes; directed networks have two types of degrees, in-degree and out-degree, which are the number of incoming and outgoing edges, respectively. Nodes with a high degree are called hubs, and these hubs tend to exert a large amount of control on the network compared with a node with few connections. For example, if a node with low degree is removed from the network, few connections will be disrupted and overall network structure will be maintained. However, if a hub node is removed, several connections will be removed along with it, resulting in fragmentation of network topology. Transcription factors are often network hubs due to coexpression with many target genes. Transcription factors within networks can also be ranked based on their hubbiness once regulatory edges describing potential protein-DNA binding interactions are included.

Neighborhood connectivity is the average connectivity of all neighbors of a given node. The clustering coefficient is a measure of how connected a node and its neighbors are to each other. An average clustering coefficient can be determined for an entire network and indicates the overall level of clustering possible for a given graph. Betweenness centrality calculates how central a node is within a network and indicates that node’s level of influence on its neighbors and the network as a whole. Likewise, “coreness” is a metric that describes the location of a node within a network, specifically how centrally it is located in the network, which can also be used to evaluate the influence of a particular node on a network.

**Hubs, Hubbiness, and H-Index**

A common feature of biological gene networks is that few nodes have a high degree (i.e., they are hubs) and many nodes have a low degree. (Networks with this property of a few, highly influential nodes, which are very different than random networks, are also described as “scale-free” networks.) Given the influence of a hub node on network topology, it makes sense that the majority of nodes within a biological network have a low degree, resulting in overall network robustness when challenged by perturbations. While hubbiness is a common metric to determine the influence of a node on its neighbors, it poorly describes the influence of that node on the network as a whole. Alternatively, the H-index (Hirsch index) is a metric that measures a node’s importance on whole network structure. (The H-index is also used as a measure of a scholar’s influence). In network analysis, H-index states that the importance of a node is dependent not only on how many neighbors it influences, but also on how influential its neighbors are (number of out-degrees). Recent studies have shown that
H-index is better than coreness and degree metrics alone for determining overall node influence on a network.

Guilt-by-Association

Network analysis can help identify the putative biological function of a gene node in the network through the above-mentioned principle of guilt-by-association. This is because the edges connecting the nodes in the network contain functional information about the relationship between those nodes. It has been found in many biological networks that nodes that are closely connected in a gene expression network are often involved in the same biochemical or biological process. Thus, predictions can be made for gene nodes of unknown function based on the function of their first neighbors and then tested via experiments in the lab.

Conclusions

Gene expression is regulated on many different levels and by many different factors. This teaching tool focuses on understanding regulation at the level of transcription, with specific focus on transcription factors. Genetic perturbation of transcription factors can reveal their role by examining genome-wide reprogramming of pathways and processes. Analysis and interpretation of transcriptome data, using microarrays or RNA-seq technology, is an important skill for researchers to address biological questions surrounding transcriptional regulation. While many different technical and analytical approaches are available, the general workflow presented in this tool is a practical way to conduct transcriptome analysis, although important nuances will arise for each experiment. Network analysis is a useful approach for organizing and analyzing -omic scale data. The basics of constructing a gene coexpression network are presented here. However, the reader is encouraged to explore the recommended reading to gain insight into advanced network analysis and hypothesis generation.

PART 2: HANDS-ON ACTIVITIES TO EXPLORE DIFFERENTIALLY EXPRESSED GENES

Here, we provide two step-by-step guides for students interested in learning about the tools used for differential gene expression analysis. Please view the accompanying slides and videos for screen shots of the various steps described. The analysis was performed using versions available in June 2018. Note that some of these steps can take an hour or more to run, so if using this tutorial in a course, be sure to evaluate the timings ahead of time.

Exercise 1. A Sample Workflow for Exploring Genome-Wide Regulation by a Transcription Factor Using Microarray Data

A good way to learn a new technique or analysis method is to do it yourself. The example presented here walks you through a typical workflow for analyzing transcriptome data obtained from single-channel microarray experiments. In this exercise, you will compare two genotypes (the wild type and the wrky1 mutant, which is identified as SALK_070989), in two conditions (with and without supplemental nitrogen). The full data set includes other conditions, but we will keep things simple by comparing just the four types of samples (with three replicates each).

In this exercise, you will use the example data set to identify genes in Arabidopsis that are influenced in some way by the transcription factor AtWRKY1, in the context of nitrogen response. A set of differentially expressed genes can be identified between control plants with wild-type WRKY1 gene function and mutant plants in which the expression of the WRKY1 gene has been knocked down. Analyzing these results opens the door for generating hypotheses about specific genes that might be direct or indirect targets of WRKY1, biological processes that are potentially regulated by the WRKY1 TF in Arabidopsis, and the genotype by environment (GxE) interaction due to nitrogen response in wild-type and mutant plants (i.e., how does AtWRKY1 influence plant response to nitrogen?).

By the end of this tutorial, you will be able to:
- Statistically identify significantly differentially expressed genes
- Cluster genes with similar expression patterns
- Perform functional analysis using GO
- Build and analyze a coexpression network

Importantly, you will be in a position to critically analyze the results you obtain in order to generate hypotheses about the regulatory function of the transcription factor of interest.

For this workflow, you will need the following software.

- Microsoft Excel
- A text editor such as Notepad, Notepad++, or TextEdit
- Multiple Experiment Viewer (tMeV): http://mev.tm4.org/#/welcome (you must download a local copy to use this tutorial; see under "Browse" on above website)
- Cytoscape: http://cytoscape.org/download.php (requires Java)
- Cytoscape Application: Expression Correlation: http://apps.cytoscape.org/apps/expressioncorrelation

   A. What are the experimental treatments?
   B. How many replicates per treatment are there?
   C. What platform was used to analyze the samples?

Download the “Series Matrix File.” This is a g-zipped file that will require a program to unzip (i.e., terminal or freeware like 7-Zip). The “Series Matrix File” is provided by the researchers who conducted the experiment. This file includes details about the experiment and data analysis and provides the normalized expression values for every probe on the microarray. Save this file; you’ll need it later.

2. Data selection and analysis. For simplicity, we only want to analyze a subset of the samples in this data set. We are interested in comparing wild-type and mutant plants in the presence and absence of nitrogen, and there are three replicates (reps) for each condition. Analysis will be performed using GEO2R, which is a straightforward tool embedded in GEO that allows for automatic statistical analysis using LIMMA.

On the GEO webpage, click the big blue button, “Analyze with GEO2R.”
A. Data selection: You can select the samples you want to analyze by defining groups as follows:
   a. Click “Define groups” and type “wrky1_control,” “wrky1_nitrogen,” “WT_control,” and “WT_nitrogen.”
   b. Now, select “COL-0 in N rich treatment” replicates 1, 2, and 3 and assign them to the group “WT_nitrogen.”
      Select “COL-0 in N absent treatment” rep 1, 2, and 3 and assign them to the group “WT_control.”
      Select “SALK_070989 in N rich treatment” rep 1, 2, and 3 and assign them to the group “wrky1_nitrogen.”
      Finally, select “SALK_070989 in N absent treatment” rep 1, 2, and 3 and assign them to “wrky1_control.”
   c. Perform LIMMA statistical analysis to calculate significant DEGs by clicking “Top 250,” which is at the bottom of the page. Use the default parameters, but familiarize yourself with the specific parameters.
   Question: Which method of correction was used for multiple comparisons?
   d. You can also get an idea if the samples were normalized correctly. Click “Value distribution” to see a box plot of a subset of the data distribution for the samples you have selected. Look to see if the medians are centered across samples or if the data are skewed. Median centered values specify that the data have been normalized and can be compared. Non-median centered plots suggest that the data are not directly comparable.
   e. Go back to the GEO2R tab and click “Save all results.”
      In the new window, select all data and copy into a text editor like Notepad in PCs or TextEdit in Macs and save as a .txt or .csv file. (Note: If using TextEdit, switch to plain text mode by selecting Format->Make Plain Text from the program menu bar.)

B. Data analysis: To begin analyzing your results, you first need to choose a cutoff for significance. It is best practice to choose a cutoff for the adjusted P value (adj.P.Val), which takes into account the multiple comparisons. This will result in a higher confidence data set that can be used for downstream analysis. In this example, the DEGs listed are those that are misregulated in response to knockdown of the TF WRKY1.
   a. Open your .txt file of GEO2R results in Excel, and sort the database on the FDR-adjusted P value (adj.P.Val) from smallest to largest values. A cutoff of < 0.01 will give you 861 probe IDs that you have confidence in differential expression. A P value of 0.01 means that you expect to reject the null hypothesis 1% of the time due to random noise in the data.
   b. Create a new sheet in your workbook to list the identities of the probes (genes) that are significantly different in their expression levels between the treatments. Call this sheet “Sig_IDs.” Copy the column of probe IDs (Column A) for P values all values under your cutoff and paste into the new sheet in Column A.
   c. For downstream analysis, including clustering and gene network, you will need to add to this sheet the normalized expression values from the Series Matrix File that you saved earlier. Remember, you only want to copy in the values for the subset of 12 samples (2 genotypes × 2 conditions × 3 replicates each). Create a new sheet that includes the probe names (in column A) and the data columns for the 12 samples of interest (GSM1978990-GSM1979001) and name it “Trimmed.” Optional: You may replace GEO identifiers with more intuitive sample names such as WT_N1, WT_N2, WT_N3, wrky1_N1, etc.
   d. Next, you need to complete the SIG_ID worksheet with the expression level values for each sample for each of the probes (that you sorted in order of significance). You will use the VLOOKUP function in Excel to pull into the SIG_ID sheet the normalized expression values for your significant probe IDs.
      i. First, in the “Sig_IDs” sheet, give a header to 12 columns (B to M) in the same order as they appear in “Trimmed.” i.e., WT_N1, WT_N2, WT_N3, wrky1_N1, etc.
      ii. Next, apply the “VLOOKUP” function (see the video tutorial if you are unfamiliar with VLOOKUP). This function instructs excel to fetch the value for a given probe from the appropriate cell in the Trimmed worksheet and paste it into the SIG_ID worksheet.
         First, you will select the probe ID name in column A of SIG_ID worksheet that you want to pull the value for. Then you will select the source file for the expression data values. Then you will indicate which column of values to pull the data from.
      iii. To do this, place your cursor in Column B, Cell 2. Type “VLOOKUP(A2, then select the entire worksheet, give a header to force
         "Trimmed", which column of values to pull the data from.)
         Note: In the formula, put the $ symbol in front of A2
         (column 3. The simplest way to do this is to copy the
         function bar change the “2” to a “3.” As you move right through
         the worksheet, you will need to tell the function which column in “Trimmed” the value of interest can
         be found. For example, for Cell D2 (WT_N2), the
         VLOOKUP function will still use value A2 as the ID to
         look up, and it will be the same sheet “Trimmed,”
         but you need to now type “4” instead of 2 to indicate
         the next value will be found in the fourth column.
   e. Finally, make a copy of “Sig_IDs” including
      the column headers, but paste just the values (use the
interactions due to combined treatments of expression. You will also be able to observe more complex sequentially in the same process would be coexpressed. The underlying assumption is that genes that act together/aligned will be grouped together and those with very different which sets of genes whose expression patterns are closely tive way to identify patterns in large-scale expression data, in genes with common expression patterns. Clustering is an effec-

3. Cluster Analysis: Clustering algorithms let you visualize genes with common expression patterns. Clustering is an effective way to identify patterns in large-scale expression data, in which sets of genes whose expression patterns are closely aligned will be grouped together and those with very different expression patterns will be placed farther apart in the output. The underlying assumption is that genes that act together/sequentially in the same process would be coexpressed.

In this example, you will be able to identify genes that are up- or downregulated in response to knockdown of the TF WRKY1 expression. You will also be able to observe more complex interactions due to combined treatments of “genotype” and “nitrogen regime.” For example, look for genes that are differentially expressed only under a particular combination of treatments such as Mutant + High Nitrogen.

Multiple Experiment Viewer (MeV) is a free tool that is commonly used for cluster analysis and visualization. To follow these steps, please download the appropriate MeV version for your computer from http://mev.tm4.org/#!/welcome. You must download a local copy to use this tutorial: see under “Browse” on website and select “Download Current Stable Version (zip).”

A. Open the downloaded MeV application and Click “File” then “Load Data.” Optional: select “Load Annotation” for Arabidopsis thaliana.

B. Use “Browse” to find the “Sig_Values.txt” file you saved in Step 2 for “Single-color Array” and click the upper-leftmost expression values. Click the “Load” button to finish.

C. In MeV, you can perform adjustments to the data to more easily visualize differences as a heat map. If you choose not to transform the data, you should still “set color scale limits” under the “display” tab to adjust the upper and lower limits based on the suggested values. To adjust the data:

a. Click Adjust Data → Gene/Row Adjustments → Normalize Genes/Rows. This will transform the loaded expression values using the mean and the SD of the row of the matrix for that value, as follows: Value = [(Value) − Mean(Row)]/[SD(Row)]

b. Click “Original Data” in the left column, and select “Expression Image.” Everything should be one color.

c. Click display → set color scale limits → Fill in the boxes with the suggested value in the “Color Range Selection” box (i.e., −2.27 for lower limit). This will reduce the number of elements below the lower limit and above the upper limit, which you want to be close to zero.

D. Next, perform clustering as follows: Analysis → Clustering → QTC. Use the default parameters in the box, → OK. QT clustering is based on maximum cluster diameter, which is the largest distance (between 0 and 1) allowed between two genes. A small cluster diameter, closer to 0, produces small, nonvariable clusters.

E. View your results by selecting “QTC - genes” from the left column. Double click “Expression Graphs” and select “All clusters.” You can also look at individual heat maps for each cluster by double-clicking “Expression Images” and clicking each cluster. Each cluster is a set of genes that share a related expression pattern, for example, upregulated by nitrogen or downregulated in the mutant.

F. Questions:

a. How many clusters did QTC return?
b. What are the obvious patterns for each cluster?
c. Why are there “unassigned genes”?    

G. To save your clusters and images:

a. Right click on one of the graphs from the “All Clusters” Expression Graph and choose “Save All Clusters.” Give the file a name such as QTC_analysis.
b. Under “File” select “Save Analysis As” and give it a name such as “QTC_analysis.”

4. GO: Now that you have identified genes that are clustered based on similarity of expression patterns across samples, you can investigate the biological significance of each cluster using gene ontology. GO is a set of structured, controlled vocabularies representing properties of gene products that provide information about their molecular function (e.g., binding activity), biological process (e.g., signal transduction), and cellular component (e.g., rough endoplasmic reticulum) domains. Performing this analysis may reveal entire biological processes that are regulated by the WRKY1 transcription factor and/or nitrogen treatment.

A. Go to agriGO website (http://bioinfo.cau.edu.cn/agriGO/) to perform GO analysis. For more information on agriGO, see Du et al. (2010); note also the Manual and FAQ links on the page):

B. Select Analysis Tool: Singular Enrichment Analysis (SEA).

C. Select the species: Supported species → Arabidopsis thaliana.

D. Copy and paste gene IDs (e.g., AT5G38530) from Cluster 1 QTC results into the “query list” box.

E. Select reference background: Suggested backgrounds → Affymetrix ATH1 Genome Array (GPL198) since this is the platform that was used for initial transcriptome analysis.

F. Select Advanced options: Statistical test method = Hypergeometric; Multi-test adjustment = Hochberg (FDR); Significance level = 0.1; Minimum number of
mapping entries = 3; Gene ontology type = Complete GO. Optional: name your job and enter your email address for notification.

G. Browse the graphical results and chart for each cluster
H. Observe and download results by selecting the “download” button and saving results as “Cluster1.GO.”
I. Repeat for each of the clusters you identified in the cluster analysis in Step 3.

Alternatively:
A. If you haven’t already generated a Sig Values sheet with the probe IDs replaced with AGI gene names, you need to do that before using Panther. You can do that using this site: http://bar.utoronto.ca/ntools/cgi-bin/ntools_agi_converter.cgi. Select “From Probeset” and “Two AGI”, then paste the probe ID list in, and replace them with the AGI names in each cluster sheet.
B. Go to the Panther Classification System website to perform Gene List Analysis: http://pantherdb.org/
C. Copy and paste gene IDs (e.g., AT5G38530) from Cluster 1 QTC results into Box 1 and choose “ID list” as list type.
D. Under “Select organism” choose Arabidopsis thaliana.
E. Under “Select analysis” choose Statistical overrepresentation test and UNCHECK “use default settings.” On the next page under “Annotation Data set” choose “GO biological process complete,” then launch analysis.
F. Export results and save file as “Cluster1.GO.” Repeat this process for all clusters from the QTC analysis. Note: For small clusters, you should deselect “Use the Bonferroni correction for multiple testing” to loosen the statistical criteria and return results. Correction for multiple testing makes more sense for large lists (see text above). Also, the default is to display results with P < 0.05 so for small lists you will want to “display all results” to see if you have results with P < 0.1 or whatever you determine is an appropriate cutoff. Always ask yourself if the results make biological sense, whether or not they pass an arbitrary P value cutoff.
G. Now, compare results of the cluster analysis with results from the GO analysis. For example, the genes in Cluster 1 are upregulated in the presence of nitrogen, but downregulated in the absence of nitrogen, regardless of plant genotype. The most statistically overrepresented GO categories for these genes include those related to nitrogen response such as “response to chitin,” which has a parent term “response to nitrogen compound and “ncRNA processing,” which has parent terms “cellular nitrogen compound metabolic process” and “nitrogen compound metabolic process.” However, there are also overrepresented terms that are not directly linked to nitrogen response, but have a biological link to cellular nitrogen content, such as “ribosome biogenesis,” which requires nitrogen containing molecules (amino acids) to build the ribosome and downstream proteins. Do the same analysis for the other clusters from QTC analysis and determine if the results make biological sense in respect to the experiment and the expression results.

5. Coexpression network construction: Networks are useful in understanding the regulatory role of a transcription factor on a whole system through construction of coexpression and/or gene regulatory networks. Network analysis is based on graph theory, where a graph is defined as a pair (V, E) where V is a set of vertices or nodes, and E is the set of edges connecting the nodes. Such graphs can be directed or undirected, in which directed edges are applied to ordered pairs of nodes as would be expected in a stepwise biological or metabolic process.

Building networks, or interactomes, helps to organize -omics scale data into simple graphs that depict relationships between genes. In a coexpression network, the edges between nodes are defined by the level of correlation for their expression across samples. Network analysis can identify features within a network such as closeness centrality and hubbiness (see lecture notes and slides for more). Cytoscape (http://www.cytoscape.org/) will be used for network construction and coexpression analysis. Start by downloading Cytoscape. (Note: Tutorial was made using version 3.6.1.)

A. Open Cytoscape and start a new session by closing the welcome box (note: Cytoscape can open slowly).
B. “Import Table,” select your “Sig_Values.txt” file. Then open the application “ExpressionCorrelation” and choose “Construct Correlation Network.” Name the table Network1.
C. The default correlation cutoff is −0.95 and 0.95, which results in more than 14,000 edges. Large networks become difficult to analyze and interpret their results. For the purpose of this tutorial, let’s decrease the size of the network to something more manageable. Go to Apps → ExpressionCorrelation → Advanced Options → Gene Network: Preview Histogram. The graph shows you the distribution of interactions at different correlation coefficients. The majority of interactions have a correlation coefficient between −1.0 and −0.7 and 0.7 and 1.0. To achieve a manageable number of interactions, decrease the low cutoff to −0.98 and increase the high cutoff to 0.98; this will result in fewer interactions (~800).
D. You can change the layout of the network to visualize it in a preferred way. For example: Layout → yFiles Layout → Organic.
E. For network analysis, go to Tools → NetworkAnalyzer → Network Analysis → Analyze Network → Treat the network as undirected.
a. Explore the various tabs in the Results Panel such as Closeness Centrality, Neighborhood Connectivity Distribution, and Node Degree Distribution. The graphs in these tabs show total network statistics. You can save these if you wish.
b. Minimize the Results Panel and look at the Table Panel for probe specific information.
c. Sort the Table Panel by Degree by clicking “Degree.” Degree is the number of connections (edges) a node. Nodes with a high degree are called hubs, and these hubs tend to exert a large amount of control on the
network compared with a node with few connections.

Question:
  i. What nodes have the highest degree?
  d. Now, sort the Table Panel by Clustering Coefficient.
  Question:
  i. What nodes have the highest clustering coefficient?
  e. Compare the results of your network analysis to those from the clustering and GO analyses. Question:
  i. What conclusions can you draw about the most influential nodes in the network and the biological processes they are involved in?

Exercise 2: A Sample Workflow for Exploring Genome-Wide Regulation by a Transcription Factor Using RNA-Sequencing Data

This tutorial will guide you through DEG analysis of an RNA-seq experiment. In this example, RNA-sequence files will be analyzed (whereas in Exercise 1, microarray samples are analyzed). In this study, RNA samples were obtained from wild-type Arabidopsis floral buds, and floral buds from plants in which the transcription factor RBE was transiently overexpressed using the dexamethasone (DEX)-inducible promoter. Much like the example in Exercise 1, this workflow can be used to identify genes that are misregulated in RBE-overexpressing plants compared with control plants with wild-type gene function. The tools that are used to analyze RNA-seq data are different than those used to analyze microarray data, but once you have generated a list of DEGs you can use the same tools for downstream analyses (cluster analysis, GO, and coexpression network construction) as described in Exercise 1.

For this workflow, you will need the following software:

- Galaxy account (note that you need to request a Galaxy account before you start the analysis)
- Microsoft Excel
- A text editor such as Notepad or Terminal


2. Find individual sequence IDs in the SRA database
   a. Obtain the list of individual RNA-seq runs using the SRP ID (SRP refers to a group of related RNA-seq runs stored in the SRA database). For example, [https://www.ncbi.nlm.nih.gov/Traces/study/?acc=SRP011435](https://www.ncbi.nlm.nih.gov/Traces/study/?acc=SRP011435). In the top right corner, click on the “Send to” link and select “Run selector”, then “Go.”
   b. From the new webpage, copy the list of Run IDs (SRR444595-SRR444602) into a text file on your computer.

3. Import data into Galaxy
   a. If you have logged out, log in to your Galaxy account ([https://usegalaxy.org/](https://usegalaxy.org/)).
   b. In the left column, scroll down to “NCBI SRA Tools” and select “Download and extract reads in FASTA/Q format from NCBI SRA.” In the main window, paste the accession numbers (starting with SRR.) one at a time into the Accession box, followed by “Execute.” Repeat for all eight runs. The process of importing RNA-seq data into Galaxy can take up to an hour due to the very large size of such files.

4. Quality control
   a. Use the Trim Galore! tool to trim low quality bases from the RNA-seq reads. This tool, which can take more than an hour to run, identifies and eliminates low-quality bases (a low quality value implies higher chance that the base called was wrong) and identifies and removes adapter bases (adapter bases in the sequencing read are technical artifacts of the sequencing process). Removal of low quality and adapter bases allows better and quicker matching of reads to the gene sequences.

5. Gene expression counts
   a. In the left column, under NGS: RNA Analysis, use the Salmon tool to convert raw read data into gene expression counts.
   b. Initiate one instance of Salmon per sequence run using the trimmed sequence set generated in Step 4, and the coding sequences set (CDS) downloaded in Step 1a.

6. Statistical test to find DEGs
   a. In the left column, under NGS: RNA Analysis, use the DESeq2 tool to identify set of differentially expressed (DE) genes.
   b. Assign a name to the “Factor,” which in this case refers to the treatment, e.g., RBE.
   c. Assign a name to “1:Factor level” as DEX (referring to the DEX treatment to induce RBE).
   d. Select Salmon output from Dex samples (SRR444595, SRR444596, SRR444597, and SRR444598).
   e. Assign a name to “2:Factor level” as Mock.
   f. Select Salmon output from Mock samples (SRR444599, SRR444600, and SRR444601).
   g. Select TPM values from the “Choice of Input data” dropdown menu.
   h. Specify the previously uploaded Arabidopsis GTF file from Step 1b (if the previously uploaded GTF file isn’t recognized, click on it in the right column, select “edit,” and then select “datatypes” and “gtf” from the dropdown menu).
   i. Initiate a DESeq2 run using the Execute button.
   j. DESeq2 output includes table with seven columns, with the FDR corrected P value in Column 7.
The workflow in Exercise 1 from Step 3 onward (clustering, GO, and coexpression networking clustering) can be followed using the DEGs obtained from Galaxy.

**RECOMMENDED READING**

(This is a representative list of sources to help the reader access a huge body of literature. We apologize in advance to those whose work is not included.)

**Introduction to Gene Regulation, Transcription, and Transcription Factors**


Differential Expression Analysis: Tools, Techniques, and Outcomes


Functional Analysis of DEGs


