Supplemental Material

CBE—Life Sciences Education

Makarevitch *et al.*

Understanding Transcriptional Response to Abiotic Stress: Large Data Analysis in the Classroom

After completing this Lab, you should be able to:

- explain the ideas of transcriptional response of organisms to stress;
- explain the principles of RNA-Seq data analysis;
- perform basic RNA-Seq data analysis using DNA subway and DESeq and identify genes differentially expressed between two samples;
- build several types of graphs to visualize RNA-Seq data.

Introduction/Background: *Cold Stress and Maize*

Maize is an important staple crop of many different cultures and countries around the world. One of the limitations of maize growth is cold temperatures early in spring, when maize seedlings germinate. Farmers in Minnesota, Iowa, North Dakota and other US states closely watch May and early June temperatures because they could greatly affect the health of corn plans and the year's corn harvest. Changing climate poses an additional concern for corn growers and breeders. Unpredictable and severe weather patterns that have become more common recently require that crop plants become more resistant to environmental stress, including cold. To develop maize plants that have high resistance to cold stress, breeders and geneticists need to better understand the mechanisms of genetic control of abiotic stress resistance.

Transcriptional Response to Cold Stress

While almost all cells in a particular organism have the same set of genes / the same genome regardless of environmental conditions, different genes are activated in response to the environment. Such differential gene activation results in the presence of various amounts of RNA molecules transcribed from certain genes. Most of the genes will likely not change the level of their expression in response to cold stress. Some genes will be down-regulated, that is the amount of RNA molecules produced from these genes will be decreased. Some of the genes will be up-regulated, that is the amount of RNA molecules produced from these genes will be increased in the plants subjected to cold stress compared to plants grown under normal conditions.

Lab Goal

The objective of this lab is to perform initial analysis of transcriptional response of maize seedlings to cold stress and identify sets of genes differentially expressed in response to cold stress.

Lab Assessment

Please submit the answers to worksheets 1 – 3. Please also submit a lab report in a form of a scientific paper describing the problem you addressed, materials and methods that were used, results and a short discussion including at least two graphs and / or tables.

Understanding Transcriptional Response to Abiotic Stress: Large Data Analysis in the Classroom

Worksheet 1. Transcriptional Response to Abiotic Stress: Primary Literature Analysis and Developing Testable Hypotheses

The questions on this worksheet will guide students to understanding the concepts of gene expression changes in response to changes in environmental conditions and will allow students to develop testable hypotheses about variation in gene expression.

B73 and Mo17 are common maize varieties (inbreds) in the USA. They are completely homozygous (inbreds) and are used to develop new varieties as well as study maize genetics. Maize genome has been sequenced (a complete sequence of maize DNA determined) for B73 inbred and many of the molecular marker maps have been developed for these two inbred lines.

The plants you see are B73 and Mo17 seedlings (14 days after planting) that were exposed to cold stress (16 hours at 5^0C) or heat stress (4 hours at 50^0C) followed by a period of recovery at room temperature.

Please read: *Dear students! For the next couple of classes you will be scientists working on an interesting biological problem. We just received the RNA-Seq data from the plants you grew last month and I do not know the answers and results of this study. You are the first! Scientists frequently start with building predictions and hypotheses. These predictions and hypotheses do not have to be correct all the time! Scientists read some literature, think about the problem, and formulate their predictions based on what they know. It is OK to be wrong!!! Hypotheses that scientists test frequently turn out to be wrong!!! Testing them, though, frequently results in better understanding of the problem and formulating new hypotheses! I am not looking for "perfectly correct predictions" for the next several questions!* ***I would like you to support your hypotheses*** *with arguments that you derive from literature, textbook ideas, common knowledge, and other sources.* Please discuss the following questions with your group. I am happy to join the discussion at any moment: just let me know when you need me.

Question 1. Please take a look at the plants labeled B73 and Mo17. Do you see any differences in the way these two lines respond to cold stress? Describe the differences and similarities you see and make a conclusion about cold resistance / susceptibility of Mo17 and B73 maize seedlings.

Question 2. Conduct literature search (PubMed is a great resource) and find **two** manuscripts describing the effects of stress exposure (not necessarily cold stress) on activation or down-regulation of gene expression in plants or animals. Briefly summarize the experimental system, main methodology, and major results of these studies (2 – 3 sentences per each study).

Question 3. Use available resources (PubMed, Internet, textbook, or your knowledge about types of genes plants have) and predict what groups of genes you would expect to be down-regulated in response to cold stress? What types of genes you would predict to show no response to stress exposure? What types of genes would likely be activated in response to stress?

Question 4. What proportion of maize genes (~40,000 predicted high confidence genes) you would expect to be expressed in maize 14 day old seedlings? What proportion of maize genes (~40,000 predicted high confidence genes) you would expect to respond to cold? What is the basis for your prediction?

Question 5. Given the different response to cold between Mo17 and B73, what differences would you predict in the level, number, or types of genes activated or down-regulated in response to cold stress exposure between these two genotypes?

RNAs from Mo17 and B73 cold-stressed, heat-stressed, and control plants were purified and sent for RNA-Seq analysis. We have three replicates of each condition in two genotypes, B73 and Mo17.

Question 6. Please discuss with your group and formulate two questions you would like to ask / hypotheses you would like to test using this data set.

Understanding Transcriptional Response to Abiotic Stress: Large Data Analysis in the Classroom

Worksheet 2. RNA-Seq Analysis: Principles

The questions on this worksheet will guide students to understanding the principles, technology and uses of RNA-Seq to characterize transcriptional responses.

Question 1. When RNA is purified from any plant or animal tissue, what three major classes of RNA will be isolated? Which of these classes will be the most abundant? If you run total RNA on a gel, what will you likely see? Why?

Question 2. What class of RNA is the most interesting one for understanding transcriptional responses? Why? What is the common feature of these RNAs?

Question 3. To determine the sequence of all mRNAs, they must be converted to double stranded DNA (cDNA). What enzyme converts RNA to cDNA and where does this enzyme come from?

Question 4. Find available resources (YouTube, Wikipedia, http://www.rnaseqforthenextgeneration.org, etc.,) and use them to draw ***your own*** scheme explaining the major principles of RNA-Seq technology and analysis. The drawings and/or explanations should include the terms: total RNA, mRNA, cDNA, RNA-Seq library, fragmentation, adapter ligation, indexing, next-generation sequencing, and differential gene expression.

Understanding Transcriptional Response to Abiotic Stress: Large Data Analysis in the Classroom

Worksheet 3. RNA-Seq Analysis: Data Quality and Initial Analysis

The questions in this exercise will guide students to understanding the structure of FastQ RNA-Seq data files and the principles of quality control (essentially, the function of FastQC), as well as the FastQC Analysis on DNA Subway Green Line. The second part of this worksheet will help students understand the approaches used to map and count short RNA reads from RNA-Seq experiments (essentially, the function of Tophat).

RNAs from Mo17 and B73 cold-stressed, heat-stressed, and control plants were purified and sent for RNA-Seq analysis. We have three replicates of each condition in two genotypes, B73 and Mo17. Sequencing of our RNA libraries was conducted at the University of Minnesota Genomics Center.

Part 1. FastQ Data Analysis and Quality Control: How Do You Know if Your Reads Are Good Enough?

The sequence reads you will receive from the sequencer will be organized in short reads (50 – 150 bp depending on the details of sequencing approaches you chose). There will be a lot of reads – usually over 10,000,000 of them! Some of the reads will be good and some of the reads will not be so good, simply because the machine that "reads" DNA can make mistakes! Sometimes, this machine "suspects" that it makes mistakes and tells you that it is not so sure about the quality of the read. So each nucleotide has a reported "goodness score" associated with it. If you look at a sample FastQ file (look here for some examples: http://en.wikipedia.org/wiki/FASTQ_format) you can see a small number of the reads that are representative of a large set of short read sequences.

Question 1 – Sample FastQ File. Please open a Sample FastQ File from your folders and look at it. *What symbol does each sequence start with? How many sequences do you see in this file? How long are the sequences? Are they all the same size? What are some other similarities / differences that you can find in these sequences? Please look at the http://en.wikipedia.org/wiki/FASTQ_format webpage and explain in your own words what are the symbols in the fourth lane of each of the sample and why are they important?*

Question 2 – All Reads Together – Sample File. The letters on the fourth line of each sequence entry refer to the quality of a particular base (in other words, how much one can trust the data from this particular read). However, quality of a particular read might be not as important as the quality of the whole data set and it is really difficult to simply look at the sequences and infer whether they are bad or good, right? So we will use a program called "FastQC" to help us out.

If you do not have your data yet, you can google ***"DNA subway"*** and login as guests.

- Find a public ***"Green Line"*** project called "Maize Abiotic Stress" and go into "Manage Data". If the analysis is completed you should be able to see the green letter "v" close to

the "Manage Data" tab. You also should be able to view the quality of the sequences of this project by clicking on "View".

- Look at two of the samples under ***Basic statistics*** and **record** the name of the sample, and the number and length of reads in the sample. Compare several samples together. *What can you say about how similar they are in the number of sequences read in each sample?*
- Look at two of the samples under ***Per sequence quality scores***. This graph shows the distribution of all the sequence reads (y axis) in the dataset relative to their mean quality score (x axis). *What can you tell about the quality of your data? Do the two graphs describing data quality tell a similar story?*

If you have started your own project in Green Line and have your data in, you should be able to click R icon and assess the quality of your own data! If you work with your data, answer the questions about all samples from your dataset.

The other graphs in QC (Quality Control) report show a summary of other properties of your data. Although they are important and could be very useful in interpreting data results, most of the sequence runs fails at least one of these properties, sometimes because of the nature of genomic sequence of the species people work with, sometimes because of the details of sample preparation. Usually data analysis could be conducted successfully regardless of these features.

Part 2. Mapping and Counting Short Reads: Where do My Reads Come from?

You just performed quality control on your RNA-Seq reads and they seem to be of good quality. You are relieved and ready for the next step. Well, remember that your goal is to count how many reads in your samples belong to all of the genes in your experimental organism under each of the experimental conditions.

Analogy to think about:

Imagine that you want to compare cars in two different states to show that drivers in California prefer small and efficient cars, while drivers from Minnesota prefer durable four-wheel vehicles. To answer this question, you drive around the streets and highways in both states and take pictures, many pictures. You come back home with your phone filled with photos and start data analysis. First, you need to determine if your photos are good. You want to make sure that each picture is of a good quality and you can match it to a particular car type. Next comes the process of matching photos you have to car makes and models. Finally, you will count how many types each of the car models is captured in your pictures.

OK, let's go back to our RNA-Seq data. Each of the reads we have is a photo of a car. You already performed the quality control and know that most of the reads you have are of good quality (pictures are sharp). We now need to match the reads to individual genes and count them up. One of the problems with matching RNA reads comes from their size: they are short and represent only a small portion of the gene. In our analogy, the photos you took would not show complete cars, but rather small portions of the cars (a roof or a bumper, etc…). Let's see how this is done.

Exercise on read mapping and counting:

Let's say we have three genes: red, green, and blue with the following sequences: red: ATGTGATCAGTACGATACGTAGGGCAT,
green: GACTGGACTAGGGCATATCGACAT, and
blue: TTTTGTTAACGTCAGATCGGAT.

Our data contains the following short reads: TTAACG, TTTGTT, TACGTA, CGACAT, GACTAG, AGGGCA, TACGAT, GATCGG, TCGGAT, GGGCAT, TTGTTC, GTACGA, GATCAG, CGTCAG, TTGTTA, and TAGGGC.

Question 3 - Mapping. Please "map" all of these short reads to the three genes: red, blue, or green. *Can you map all of the reads? Were there reads that could not be mapped? Where do you think "unmappable reads" could come from in a real RNA-Seq experiment? Were their reads that could not be uniquely mapped, that is mapped to only one place in the genome? What genomic features are likely to produce RNA reads that could not be uniquely mapped?*

Question 4 – Counts. OK, we are done with the difficult work of mapping and now simply need to count how many times each gene was "found" among the short reads. *Please complete the following table to record the number of uniquely mapped reads corresponding to each gene in our small sample:*

Gene	**Sample 1**
Red	
Green	
Blue	
Total	

Question 5 –Normalization. Let's say we analyzed the second sample (Sample 2) and counted the reads corresponding to three genes in a similar fashion. Here is the data that were received:

Gene	**Sample 2**
Red	40
Green	3
Blue	87
Total	130

What conclusions can you make about gene expression (the difference in number of reads) in the samples 1 and 2? Can you use read counts directly to make this conclusion? Why, or why not?

You can complete the analysis of your data using Green Line. However, it takes a long time and we will not be able to complete it in class. Therefore, I completed some of these steps for you and compiled a csv file that contains raw (not normalized) read counts for all samples in our experiment (three replicates of control, cold-stressed, and heat-stressed B73 and Mo17 plants).

Understanding Transcriptional Response to Abiotic Stress: Large Data Analysis in the Classroom

Worksheet 4. Data Analysis: Finding Differentially Expressed Genes

Find a file "Data for RNA-Seq Lab" in your folder and download it to your computers. Please look back at the questions/hypotheses you would like to test (***Questions 4 – 6 from the Worksheet 1***) and identify samples that you need for your analysis. Delete all other samples from your file and save it as a ***csv*** file into a working folder where R can access it.

Now it is time to compare stress and control conditions and identify genes that change their expression level in response to stress. We can simply calculate average values for each gene's stress and control conditions and compare the ratio of these values. However, this approach will not let us estimate how robust the differences we find are and what the level of statistical significance of these changes is. We need to perform a statistical test to compare variation between stress and control conditions to variation between replicates. Program DEseq that could be run in R environment will help us in this endeavor!

DEseq Analysis: What to Do

1. We need to download, install, and load an R package called *DESeq*.

- In R type in the following:

```
source("http://bioconductor.org/biocLite.R")
biocLite("DESeq")
```

- Go to Packages / Load packages and follow prompts for loading a *DESeq* package.
- Here is a DEseq tutorial that you can always go to in case you have questions and something is not clear. http://www.bioinformaticslaboratory.nl/twikidata/pub/Education/BioinformaticsII-Seq/DESeq-tutorial/DESeq-tutorial.pdf

2. Next step is to load the csv file you prepared for analysis.

```
>DataName <- read.csv("FileName.csv",row.names =1, header=TRUE)
>head(DataName)
```

Here Filename is your file and DataName is whatever ID you want to call the table inside R that will hold your data. **Please use the raw counts file for DESeq Analysis.** Remember to make sure that the files you want to read into R are located in working directory. You may want to use commands

```
>getwd(),
>setwd(), and
>dir()
```

to send R into corresponding directory (folder).

3. Now we need to tell R what conditions are present in our data and what conditions it needs to compare. Each of your columns should be designated as either "condition1" or "condition2", where condition1 could be "control" and conditions2 could be "cold". All replicates of the same condition would have the same designation in this case. If you have questions please show me your design at this stage to avoid potential problems later on. Here are the commands to get you there but remember that the specific number of columns will depend on the composition of your

file. Simple comparison of cold and control conditions in B73 would have 6 data columns: 3 replicates of "control" and 3 replicates of "cold". Arguments under libType describe the type of the library that will be analyzed, so please leave "paired-end" there and if you are curios I will be happy to explain what this means.

```
>Design <- data.frame(row.names = colnames(DataName),condition =
c("control", "control", "control", "cold", "cold", "cold"),
libType = c("paired-end","paired-end","paired-end","paired-end", "paired-
end", "paired-end" ))
```

You can check the design of your experiment by looking it up:

```
>Design
```

Please make sure that it is correct and all of your columns have correct designations.

4. We are almost there☺

```
>conditions <- Design$condition[]
>library( "DESeq" )
>cds <- newCountDataSet( DataName, conditions )
```

These lines allow us to create a list of all conditions that we will use. It is essentially the same as the list of conditions you specified in your Design (see above). They also prepare the dataset for further analysis.

5. The next step is to normalize the library (remember our lecture exercise when we discussed that some samples could be sequenced deeper and simply have more reads for ALL genes) to the depth of the sequencing.

```
>cds <- estimateSizeFactors( cds )
>sizeFactors( cds )
>cds <- estimateDispersions( cds )
```

Question: What can you tell about the depth of sequencing for the samples you are analyzing? Were they sequenced similarly or some samples have a lot more reads that others? If we divide each column of the count table by the size factor for this column, the count values are brought to a common scale, making them comparable.

6. Finally, our data are ready for the test of differential expression!!! In this command "condition1" and "condition2" should be conditions you will compare and they should match the conditions from your design (see above). This command may take a while to run but it should give you the list of all genes with the corresponding average values for conditions you compare (MeanA and MeanB), fold change (MeanB/MeanA) and statistical measure, padj-value, for the significance of expression changes.

```
>res <- nbinomTest( cds, "condition1", "condition2")
```

7. You can check the results by looking at the beginning of the list by

```
>head (res)
```

8. Let's also select the genes that are strongly significant (padj-value of <0.01)

```
>resSig <- res[ res$padj < 0.01, ]
>resSig <- na.omit (resSig)
```

9. Finally, let's separate the genes to the genes that are up-regulated under condition 2 and genes that are down-regulated under condition 2. We will also require for these differentially expressed genes to have over 2-fold change (either up or down).

```
>resUp <- resSig[ (resSig$foldChange > 2) , ]

>resDown <- resSig[ (resSig$foldChange <0.5) , ]
```

Question: Can you explain why this filtering is useful? What other criteria would you use to find the most interesting genes responding to stress?

10. Last step: let's record our data into a file that could be opened in Excel for future analysis.

```
>write.table( resUp, file="resultsUp.txt" )
>write.table( resDown, file="resultsDown.txt" )
```

11. You can now look at the lists of your differentially expressed genes and further analyze them after importing these txt files in Excel, for example.

What is next... You can already answer some of the questions from the Worksheet 1 and test the predictions you made. To look at your data you might consider various graphs and visualization methods (see Worksheet 5 for some examples). Here are some suggestions of things you might want to look at:

- Compare the differences between replicates of the same sample and between different samples. What can you tell about the quality of your data?
- Filter out genes that have very low expression levels or genes. What can you tell about expression levels of the genes differentially expressed under stress conditions? Do the genes responding to stress tend to be low expressed genes or highly expressed genes?
- Count differentially expressed genes affected by various stress conditions. What can you tell about the proportion of differentially expressed genes relative to all genes or to all expressed genes?
- How similar and how different are the lists of differentially expressed genes for various conditions?
- What are the differentially expressed genes that you found?
- **Other ideas? Happy playing!!! I am around and eager to help you at any point.**

Understanding Transcriptional Response to Abiotic Stress: Large Data Analysis in the Classroom

Worksheet 5. Data Visualization: Common Types of Graphs Used to Show RNA-Seq Data

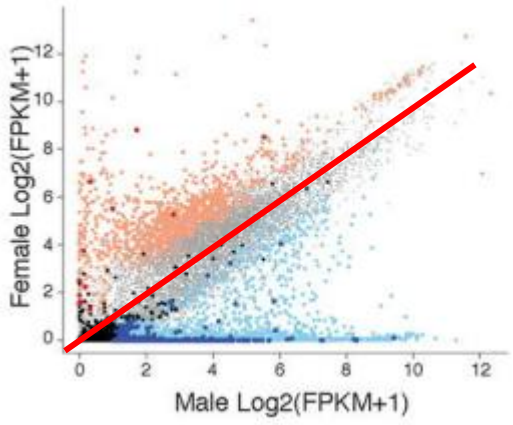
Most manuscripts that include RNA-Seq approaches to analyzing gene expression contain several different types of graphs: scatter plots, density plots, heat maps, Venn diagrams, etc. To truly understand RNA-Seq and learn to use these approaches, we need to learn how to analyze these graphs and how to visualize our own data using these graphs. In this worksheet you will explore several major types of graphs and will learn how to build them in R using your own data.

Scatter Plots

Scatter plots are frequently used to show relationships between two variables. For example, we could compare the values for expression ratios between control and stress samples or the values for absolute expression level under control condition and the expression ratio between control and stress samples.

Example: *Comparing gene expression in Drosophila males and females.* The following figure is Fig. 3b from manuscript titled *"The Developmental Transcriptome of Drosophila melanogaster"* (Graveley et al., 2010) published in Nature in 2010.

This figure shows the relationships between expression levels of all *Drosophila* genes in females and males. Each dot in this graph corresponds to a gene. Expression levels of all genes in females are plotted along y axis, and expression levels of all genes in males are plotted along x axis. The genes that show identical expression in females and males are expected to fall on a diagonal line (shown in red). *Note that both axes use logarithmic scale* to allow for better visualization of genes with low expression levels. This approach allows to better "spread out" data points with low and similar values. Note also that you can "color" dots based on some values. On this figure, the genes that show higher expression in females are colored light red, while the genes that show higher expression in males are colored light blue. There are some other groups of genes colored dark red and dark blue but we will ignore these genes for now.



Question 1. Explain what conclusion could be made based on this graph about expression of genes in *Drosophila* males and females.

Question 2. Write a figure legend that would explain this figure. Think about what information should be included in the manuscript along with this figure for a reader to understand it's meaning. Next, find the original manuscript online and compare the figure legend you wrote to the one that was included in the original paper. What did you miss? What did the authors miss?

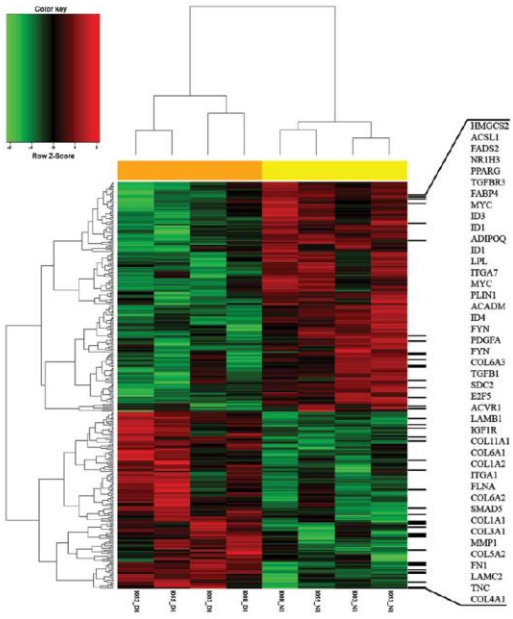
Question 3. Brainstorm the types of RNA-Seq- related questions that could be answered with the scatter plots similar to the one described above. Think about your own data set and formulate at least two questions you might want to ask. Indicate what variables you will plot on x and y axes and what could be the basis for coloring the dots in your graph.

Question 4. Use your data sets and the instructions to construct graphs and answer two questions you formulated previously.

Heat Maps

Heat maps are graphical representations of data where the individual values are contained in a matrix and each value is represented as a color (intensity of heat). They are frequently used to show relationships between different samples (for example, experimental and control conditions, multiple replicates of various samples, etc…) They are also useful to visualize the patterns of gene expression changes and to discover groups of genes with similar expression patterns.

Example: *Molecular signatures of basal cell carcinoma.* The following figure is a Fig. 1 from manuscript titled *"Molecular signatures of basal cell carcinoma susceptibility and pathogenesis: A genomic approach"* (Heller et al., 2013) published in International Journal of Oncology in 2013.



This heat map shows the relationships between tissue samples and gene expression: Tissue samples are represented by columns, each row represents a single gene found to be differentially expressed (or expressed at significantly different level) between two types of tissue samples. Colors from green to red represent the relative level of gene expression from very low (green) to very high (red). If the gene is shown in black in a particular sample, its expression does not significantly change between tissue samples. If the gene is shown in red in a particular sample, its expression level in that sample is significantly higher than in other samples. If the gene is shown in green in a particular sample, its expression level in that sample is significantly lower than in other samples. Hierarchical clustering used to construct the heat map allows to cluster (or combine) tissue samples and genes based on how similar their expression patterns are. Here, all the samples that come from disease (lesional) samples group together (shown by an orange bar) and all control samples group together (shown by a yellow bar.) Two large groups of genes are also easily observed on this figure: genes that are up-regulated (red), or expressed at higher levels, in disease samples (lower half or so of the genes) and genes that are down-regulated (green), or expressed at lower levels, in disease samples (higher half of the genes).

Question 1. Explain what conclusion could be made based on this graph about gene expression and similarity of samples derived from basal cell carcinoma and healthy tissues.

Question 2. Write a figure legend that would explain this figure. Think about what information should be included in the manuscript along with this figure for a reader to understand its meaning. Find the original manuscript online and compare the figure legend you wrote to the one that was included in the original paper. What did you miss? What did the authors miss?

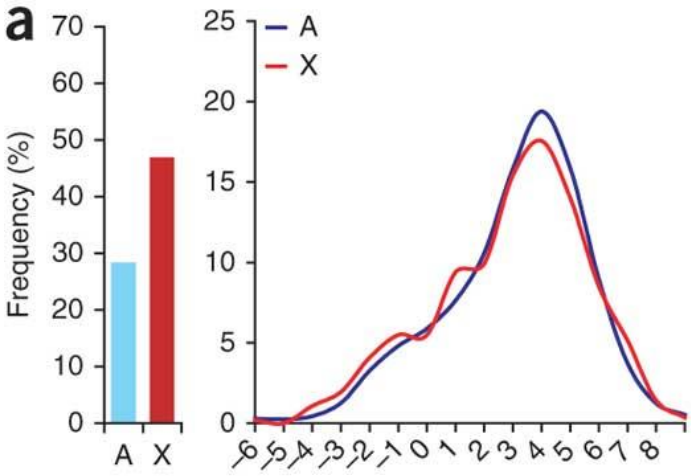
Question 3. Brainstorm the types of RNA-Seq- related questions that could be answered with the heat maps similar to the one described above. Think about your own data set and formulate at least two questions you might want to ask.

Question 4. Use your data sets and the instructions to construct a heat map clustering figure and answer two questions you formulated previously.

<u>Density Plots</u>

Frequently, it is interesting to see the distribution of data points in the data along a particular variable. For example, one might be interested in asking: What are the general profiles of gene expression levels in my sample? Or how different are the general profiles of gene expression levels under control and experimental conditions?

Example: *Comparing gene expression levels for X-linked and autosomal genes.* The following figure is a Fig2a from manuscript titled *"Evidence for compensatory upregulation of expressed X-linked genes in mammals, Caenorhabditis elegans and Drosophila melanogaster"* (Deng et al., 2011) published in Nature Genetics in 2011. This figure compares the distribution of gene expression levels for all genes located on X chromosome and all autosomal (A) genes in mouse tissues. The X axis represents the RPKM values for gene expression in logarithmic scale. Genes are grouped into bins/groups relative to the expression level. For example, all genes that have expression level between 0 and 1 RPKMs would be grouped together. Y axis shows the proportion (in %) that the genes of a certain bin represent. In essence, a density plot is a "smoothed" version of a histogram / bar graph that shows distribution of data along one variable. Since some genes are not expressed (RPKM values are equal to 0) and logarithm of 0 is not defined, the proportion of these non-expressed genes is compared separately in a bar graph on the left.



Question 1. Explain what conclusion could be made based on this graph about gene expression for genes located on X chromosome and autosomes.

Question 2. Write a figure legend that explains this figure. Think about information should be included in the manuscript along with this figure for a reader to understand its meaning. Find the original manuscript online and compare the figure legend you wrote to the one that was included in the original paper. What did you miss? What did the authors miss?

Question 3. Brainstorm the types of RNA-Seq- related questions that could be answered with the density plots similar to the one described above. Think about your own data set and formulate at least two questions you might want to ask.

Question 4. Use your data sets and the instructions to construct a density plot figure and answer two questions you formulated previously.

Histograms

Histograms essentially are somewhere in between of vertical bar graphs and density plots; a graphical representation of the distribution of data and provide a rough estimate of the density of the data. To build histograms, the data are divided between several bins/groups and the proportion of data points that fall within each bin is calculated.

Example: *Understanding the function of DNA methylation.* The following figure is a Fig3A from manuscript titled *"Function and Evolution of DNA Methylation in Nasonia vitripennis"* (Wang et al., 2013) published in PLoS Genetics in 2013. This figure compares the distribution of gene expression levels for genes that contain high levels of DNA modification called DNA methylation and genes that are free of DNA methylation. The X axis represents the RPKM values for gene expression in logarithmic scale. Genes are grouped into bins/groups relative to the expression level. Genes with similar expression levels are grouped together into the same bin and the number of genes in each bin is graphed. Y axis shows the number of genes (gene counts) in a certain bin. The gene counts could be represented in a form of proportion or a percent from the overall number of genes.

Question 1. Explain what conclusion could be made based on this graph about gene expression for genes carrying DNA methylation marks and genes that are free of DNA methylation marks.

Question 2. Write a figure legend that explains this figure. Think about information should be included in the manuscript along with this figure for a reader to understand its meaning. Find the original manuscript online and compare the figure legend you wrote to the one that was included in the original paper. What did you miss? What did the authors miss?

Question 3. Brainstorm the types of RNA-Seq- related questions that could be answered with the histograms similar to the one described above. Think about your own data set and formulate at least two questions you might want to ask.

Question 4. Use your data sets and the instructions to construct a histogram figure and answer two questions you formulated previously.

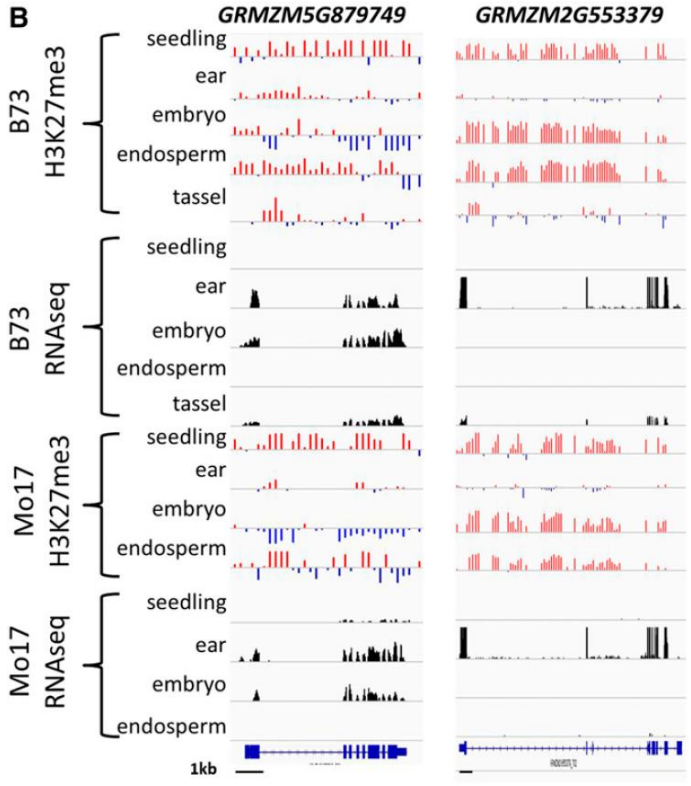
Venn Diagrams

Venn diagrams are frequently used to compare and show overlap and differences between sets. For example, assume one knows the number of genes up-regulated in one experimental condition, relative to a different condition, and wants to know how many of those genes are up-regulated either in both, or in just one or the other, a Venn diagram can be constructed to show that relationship. Venn diagrams are not limited to just 2 sets and can be used to compare 3 or more. To build Venn diagrams, one needs to know the number of genes/objects in each of the lists/sets and the number of genes/objects overlapping between the two sets. Refer to "How to Combine Gene Lists in R" to prepare your dataset for building a Venn diagram.

Genome Views

Various genome browsers / genome viewers are freely available for researchers to visualize the positions and orientations of various genome features, such as genes, transposable elements, levels of gene expression, DNA methylation, and many more (for example, see IGV, http://www.broadinstitute.org/igv/; Robinson et al., 2011). Some of these browsers allow researchers to import their own RNA-Seq data which is then mapped to the genome enabling visualization of this data in relation to the reference genome, or a Genome view, and published data tracks from other research groups.

Example: *Mapping variation of H3K27me3 histone modifications and their role in controlling gene expression.* The figure on the right is a Fig3B from manuscript titled *"Genomic distribution of maize facultative heterochromatin marked by trimethylation of H3K27"* (Makarevitch et al., 2013) published in Plant Cell in 2013. -This figure focuses on two genes, GRMZM5G879749 and GRMZM2G553379 comparing the distribution of a specific histone modification *H3K27me3*, (trimethylation of Lysine 27 on Histone H3 tails), on these genes in five tissues (seedling, ear, embryo, endosperm, and tassel) in two maize genotypes (B73 and Mo17). It also displays the gene expression level of these genes in the same tissues and genotypes as determined by RNA-Seq. So concentrate on one gene at a time. Each row / line refers to one tissue / genotype combination. Red and blue



bars show the levels of *H3K27me3* methylation (higher levels are shown in red and lower levels are shown in blue). The height of the bar is proportional to the degree of methylation (higher bars correspond to higher methylation levels). Black bars represent the levels of gene expression: you can think of them as RPKM levels or read counts mapped to specific gene regions. High black bars correspond to highly expressed genes, while low or no black bars correspond to low gene expression. Note that the gene models shown on the bottom in blue contain exons (rectangular shapes) and introns (lines in between exons). Keep in mind that RNA-seq reads only correspond to exons.

Question 1. Explain what conclusion could be made based on this graph about the developmental / tissue variation of H3K27me3 histone modifications and their role in regulating gene expression.

Question 2. Write a figure legend that explains this figure. Think about information that should be included in the paper along with this figure for a reader to understand its meaning. Find the original manuscript online and compare the figure legend you wrote to the one that was included in the original paper. What did you miss? What did the authors miss?

Question 3. Brainstorm the types of RNA-Seq- related questions that could be answered with the genome view graphs similar to the one described above. Think about your own data set and formulate at least two questions you might want to ask. What level of magnification (one gene, a region of several genes, the whole chromosome) would you look at the answer these questions?

References

Deng X, Hiatt JB, Nguyen DK, Ercan S, Sturgill D, Hillier LW, Schlesinger F, Davis CA, Reinke VJ, Gingeras TR, Shendure J, Waterston RH, Oliver B, Lieb JD, Disteche CM (2011) Evidence for compensatory upregulation of expressed X-linked genes in mammals, *Caenorhabditis elegans* and *Drosophila melanogaster*. *Nat Genet.* 43:1179-1185.

Graveley BR, Brooks AN, Carlson JW, Duff MO, Landolin, JM, et al. (2011) The Developmental Transcriptome of Drosophila melanogaster. *Nature.* 471: 473–479.

Heller ER, Gor A, Wang D, Hu Q, Lucchese A, Kanduc D, Katdare M, Liu S, Sinha AA (2013) Molecular signatures of basal cell carcinoma susceptibility and pathogenesis: a genomic approach. *Int J Oncol.* 42:583-596

Makarevitch I, Eichten SR, Briskine R, Waters AJ, Danilevskaya ON, Meeley RB, Myers CL, Vaughn MW, Springer NM (2013) Genomic Distribution of Maize Facultative Heterochromatin Marked by Trimethylation of H3K27. *Plant Cell.* 25: 780–793

Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP (2011) Integrative Genomics Viewer. *Nature Biotechnology* 29:24–26.

Wang X, Wheeler D, Avery A, Rago A, Choi JH, Colbourne JK, Clark AG, Werren JH (2013) Function and evolution of DNA methylation in *Nasonia vitripennis*. *PLoS Genet.* 9:e1003872.

How to Make Graphs in R

Constructing Scatter Plots in R

1. This line sets working directory to the folder where your file is located.

>setwd ("../")

Read the file that contains your data into R. If your file is a comma delimited text file *myfile.csv*, you can use the following command:

>mydata <- read.table ("mydata.csv", header=TRUE, sep=",")

The same command will read a comma delimited text file *myfile.txt*:

>mydata <- read.table ("mydata.txt", header=TRUE, sep=",")

The attribute **header=TRUE** indicates that the first row contains column names. The attribute **sep** describes the symbol separating the columns. If **sep = ""** (the default for **read.table**) the separator is 'white space', that is one or more spaces, tabs, newlines or carriage returns.

It is usually a smart idea to make sure that the file you read in is a correct file and it was read in correctly. One of the ways to do it is using commands

>head (mydata) to check the first six lines of the file and a header row, or

>dim (mydata) to check the dimensions of the data

2. Let's say the columns you want to plot are named ***ColumnX*** and ***ColumnY***. Now we are ready to plot the values:

>plot (mydata$ColumnX, mydata$ColumnY, main = "Title", xlab="ColumnX", ylab = "ColumnY")

The attributes **xlab** and **ylab** are used to provide labels for axes, while the attribute **main** is used to provide the title for the chart. **A word of caution: R sometimes does not "like" the quotation marks when copied from Word. I am not sure what the reason is but it might be a smart idea to retype quotation marks in R if it gives you an error of "unexpected symbols".**

3. We can also generate a linear regression model of the two variables with the **lm** function, and then draw a trend line with **abline**.

> abline(lm(mydata$ColumnY ~mydata$ColumnX))

4. Remember that plotting logarithmic values might be useful for the data you have. We can do it by using attribute **log** in the **plot** function. **log="xy"** is useful to convert values for both x and y into logarithmic scale. If only one of the axes is to be converted, you can use **log="x"** or **log="y"**.

>plot (mydata$ColumnX, mydata$ColumnY, main= "Title", xlab="ColumnX", ylab = "ColumnY", log="xy")

The program will give you warnings because some of the RPKM values we have are equal to 0 and log (0) is not defined. To overcome this problem, we may want to add a small value (0.01, for example) to all of the RPKM values before graphing:

>plot (mydata$Mean_Control_B73+0.01, mydata$Mean_Cold_B73+0.01, xlab="ColumnX", ylab = "ColumnY", log="xy")

5. Finally, remember we wanted to color the genes according to some other parameters. For example, we could color all the genes that are up-regulated under our experimental condition in red, the genes down-regulated under our experimental condition in blue, while color all genes that do not change expression in black.

Here is how you can do it. **GeneColor** is a new column that will be created in our data table (**mydata**). **mydata$GeneColor** is a column in **mydata** that will carry information about the color we would like to color the genes. This code creates an additional column in our table called **GeneColor**, assigns an initial value "Black" to all of the genes, checks for the ratio between experimental and control conditions (recorded in Columns Y and X) and the q_value (statistical significance of differences) and assigns appropriate "Color" to each gene. If the expression in experimental condition is increased more than 2 fold and the difference is statistically significant (q_value <0.001), the gene is called "Up-regulated" and colored red; if the expression in experimental condition is increased less than 0.5 fold (in essence decreased more than 2 fold) and statistically significant, the gene is called "Down-regulated" and colored blue; all the other genes remain colored black.

Note: If your data table already contains the column with the expression ratio between two conditions, you could use that column instead of calculating the ratio here. If your data

table does not contain q values for the gene expression, you can use only the expression ratio for assigning color to the genes.

>mydata$GeneColor<- NA

>mydata$GeneColor<- "black"

>mydata$GeneColor[(mydata$ColumnY / (mydata$ColumnX+0.01) >=2) & (mydata$q_value <0.001)]<-"red"

>mydata$GeneColor[(mydata$ColumnY / (mydata$ColumnX+0.01) <=0.5) & (mydata$q_value <0.001)]<-"blue"

Well, now as our gene colors are assigned, we are ready to make our plot:

>plot (mydata$Mean_Control_B73+0.01, mydata$Mean_Cold_B73+0.01, xlab="ColumnX", ylab = "ColumnY", log="xy", col=mydata$GeneColor)

Constructing Heat Maps in R

1. Make the file that contains all of your differentially expressed genes and have columns (RPKM values) for all of the samples you want to compare (replicates or average values are fine). Note: you do not want to keep the number of genes over ~15,000.

2. Let's do it:

>data <- read.csv("Query1.txt"), where Query1.txt is your comma delimited file

>head (data), to make sure your data look fine

>row.names(data) <- data$X, to record the gene names in case you would need them later

>dataM <- data[,2:19], this selects all of the columns that has data (numbers, not words); if your data has more or less columns you may want to modify the number "19"

>data_mat <- data.matrix(dataM), this converts your data into a matrix form

>heatmap.2 (data_mat, col=greenred, scale="row", density.info="none", trace="none", main ="Your Favorite Title"), this command makes your heat map. Note: it might take several minutes, do not panic☺

Do not forget to copy the picture somewhere…

3. What does it mean? Look at how your samples (columns) cluster and make some conclusions.

Constructing Density Plots in R

1. This line sets working directory to the folder where your file is located.

```
>setwd ("../")
```

Read the file that contains your data into R. If your file is a comma delimited text file *myfile.csv*, you can use the following command:

```
>mydata <- read.table ("mydata.csv", header=TRUE,   sep=",")
```

The same command will read a comma delimited text file *myfile.txt*:

```
>mydata <- read.table ("mydata.txt", header=TRUE,   sep=",")
```

It is usually a smart idea to make sure that the file you read in is a correct file and it was read in correctly. One of the ways to do it is using commands

>head (mydata) to check the first six lines of the file and a header row, or

>dim (mydata) to check the dimensions of the data

2. Let's say the variable we want to use for plotting distributions is stored in **ColumnX** of our data table **mydata**. Making density plots are actually not that difficult. We will start with calculating the density data:

```
>d<- density (mydata$ColumnX)
```

Now we can plot the data:

```
>plot (d)
```

What you will likely see is that most of the genes are expressed at low levels and the graph does not provide sufficient resolution to see the distribution well. One way around is to convert gene expression values to log scale prior to calculating density:

```
>d<-density(log10(mydata$ColumnX))
```

We can plot density distributions for several samples on the same plot by specifying different color for different lines. First, let's calculate densities for two different columns / variables:

```
>dX<-density(log10(mydata$ColumnX))
>dY<-density(log10(mydata$ColumnY))
```

Now let's plot both lines together:

```
>plot (dX, col = "red", main = "Title")
>lines (dY, col = "blue")
```

3. Finally, we could calculate how many genes in our samples are not expressed (have RPKM equal to 0) since these genes are not shown on density distributions when log scale is used:

```
>nrow (subset (mydata, mydata$ColumnX == 0))
```

Constructing Histograms in R

1. This line sets working directory to the folder where your file is located.

```
>setwd ("../")
```

Read the file that contains your data into R. If your file is a comma delimited text file *myfile.csv*, you can use the following command:

```
>mydata <- read.table ("mydata.csv", header=TRUE,  sep=",")
```

The same command will read a comma delimited text file *myfile.txt*:

```
>mydata <- read.table ("mydata.txt", header=TRUE,  sep=",")
```

It is usually a smart idea to make sure that the file you read in is a correct file and it was read in correctly. One of the ways to do it is using commands

>head (mydata) to check the first six lines of the file and a header row, or

>dim (mydata) to check the dimensions of the data

2. Let's say the variable we want to use for plotting distributions is stored in **ColumnX** of our data table. Let's filter our data to get rid of the genes with RPKMs equal to 0 (we will use log scale and logarithm of 0 is not defined).

```
>mydata<-subset(mydata, mydata$ColumnX>0)
```

Here the data were converted to logarithmic scale prior to plotting.

```
>hist(log10(mydata$ColumnX))
```

3. There are a number of things that R does by default in creating this histogram, one of the important ones is the number of breaks, or bins, it uses to group the data. One of the ways to see how R does it is as follows:

```
>histinfo <- hist(log10(mydata$ColumnX))
>histinfo
```

The first output line ($breaks) shows the breakpoints between the bins that were used. R chooses how to bin your data for you by default using an algorithm, but if you want coarser or finer groups you can do the following:

> hist(log10(mydata$ColumnX), breaks=5), where the number roughly determines the number of bins. The bin numbers do not correspond exactly to the number you put in, because of the way R runs its algorithm to break up the data but it gives you generally what you want.

The default name of the Y axis is "frequency" and the actual number of genes corresponding to each of the bins is shown. These values could be easily converted to density (proportion) by:

> hist(log10(mydata$ColumnX), freq=FALSE)

Now it is time to improve the outlook of our histogram by adjusting x-axis, y-axis, axis labels, title, and color like this:

> hist(log10(mydata$ColumnX), freq=FALSE, xlab = "ColumnX", main = "Cool Title", col="your_favorite_color")

If desired, the limits could be set for x and y axes by using parameters **xlim=c()** and **ylim=c()**.

Finally, we can add a nice normal distribution curve to this plot using the **curve()** function, in which we can specify a normal density function with mean and standard deviation that is equal to the mean and standard deviation of my data, and add this to the previous plot with a red color and a line width of 2. You can play around with these options to get the kind of line you want:

>curve(dnorm(x, mean=mean(mydata$ColumnX), sd=sd(mydata$ColumnX)), add=TRUE, col="red", lwd=2)

Combining Gene Lists in R

RNA-Seq analysis frequently generates lists of differentially expressed genes that need to be compared with each other. For example, comparing gene expression in two mutants relative to the wild type will generate two gene lists: a list of differentially expressed genes affected by

mutant 1 and a list of differentially expressed genes affected by mutant 2. In another example, lists of differentially expressed genes affected by heat or cold stress conditions could be generated. It may be interesting for a researcher to compare two lists to figure out how many genes thee lists have in common or how many genes are specific for one of the lists. This task could be done in various ways. This is a description of a relatively straightforward way to compare the gene lists in R.

1. Let's say we have two data tables (csv files named 'Trial Set1.csv' and 'Trial Set2.csv') that have several columns each. The first data table lists all genes affected by condition 1, while the second data table lists all genes affected by condition 2. Both data tables have a column 'GeneID' and several additional columns, let's say, 'log2(condition/control)', 'expression_condition', 'expression_stress', and 'q_value'. It is important that the column you will use to combine lists is named the same in both tables ('GeneID' in this example).

Let's read both of the files into R (remember that data_trial_1 and data_trial_2 are simply names of the variables where your data will be stored, you can name them whatever you like):

```
> data_trial_1 <- read.csv ("Trial Set1.csv")
> data_trial_2 <- read.csv ("Trial Set2.csv")
```

2. Now, let's combine both lists to produce a set of genes that are common between two lists. Remember that "GeneID" is the name of the column that will be used to join the tables.

```
>merged_list <- merge (data_trial_1, data_trial_2, by="GeneID")
```

Let's take a look what we got:

```
>merged_list
```

We have a table that has all of the columns combined: the columns from data_trial_1 and columns from data_trial_2. Note, that if the columns had the same name, the '.x' and '.y' were added to their names.

The list we produced is a result of "inner join", when all genes present in both lists are now present in the merged table. If there are elements in the common column of one table, but not the other, that partial data will not be included in the merged table. To include all rows, set **all=TRUE**. To include all rows from the first table, but not unmatched rows from the second, set **all.x=TRUE**; (the cells from columns in the unmatched row of the second table will be set to NA). (**all.y=TRUE** is also legitimate).

```
>merged_list_all <- merge (data_trial_1, data_trial_2, by="GeneID", all = TRUE)
>merged_list_all
```

3. OK, we have the list now. Let's rename the columns before we forgot what x and y are:

```
>names(merged_list) [names(merged_list) == "q_value.x"] <-
"q_value_condition1"
>merged_list
```

Here 'q_value.x' is a name of the column you would like to rename and 'q_value_condition1' is a new name for this column. You can also specify the index of the column you would like to rename:

```
>names(merged_list) [3] <- "q_value_condition1"
```

Here '3' is an index (order number) of the column you would like to rename and 'q_value_condition1' is a new name for this column.

4. Let's say we want to know how many genes are shared between two lists:

```
>nrow (merged_list)
```

You can now figure out how many genes are specific for the first list

```
>nrow (data_trial_1) – nrow (merged_list)
```

It appears you have everything you need to construct a Venn diagram comparing two lists!

5. Frequently, you run into a situation when one table contains a lot of different characteristics for your gene (for example, functional annotation and genomic location) and another table

contains a list of interesting (let's say differentially expressed) genes. In this case you would need to read both files into R and merge them together. You will likely need to make an outer join (using **all.x = TRUE**, see above)

Understanding Transcriptional Response to Abiotic Stress: Large Data Analysis in the Classroom

Content Assessment Test

1. Which of the following human cells contains a gene that specifies eye color?
   a) Cells in the eye.
   b) Cells in the heart.
   c) Gametes (sperm and egg).
   d) Cells in the eye and gametes.
   e) All of the above.

2. What causes maize root cells and maize leaf cells to look different?
   a) These cells have different sets of genes present in their genomes.
   b) These cells have different sets of active/transcribed genes.
   c) These cells activate the same genes; however, they receive and accumulate different molecules
      from the environment
   d) a and b
   e) a, b, and c

3. What effect does the environmental influence, for example, cold stress, have on genes of the organism?
   a) Multiple mutations are introduced to genes changing the effectiveness of gene function.
   b) Most genes do not change their structure and activity in response to cold stress.
   c) Many genes are activated/transcribed at higher level in response to cold stress exposure.
   d) Many genes are either repressed /transcribed at lower level or activated/transcribed at higher level in response to cold stress.

4. How many genes do most of the eukaryotic species have?
   a) less than 100
   b) 100 – 1,000
   c) 1,000 – 10,000
   d) 10,000 – 50,000
   e) more than 100,000

5. What do scientists mean when they talk about gene expression levels?
   a) relative number of gene copies present in a sample
   b) relative amount of RNA molecules transcribed from a particular gene
   c) actual number of gene copies present in each cell
   d) activity level of a protein coded by a particular gene that is measured by a particular assay

6. An animal is exposed to a treatment that results in chromatin changes such that the region around gene A is less condensed while the region around gene B is more condensed. What is an unlikely outcome?

a) Gene A is more active and gene B is silenced after the treatment.
b) Protein A is more abundant and protein B is less abundant after treatment.
c) RNA-seq shows an increase in A and B mRNAs after the treatment.
d) Proteomics shows a decrease in protein B after treatment.
e) RNA-seq shows a decrease in B mRNA after the treatment.

7. For which of the following steps is a statistical analysis needed for the evaluation of RNA-seq data?

a) To determine if the numbers of sequence reads aligning to a particular gene are significantly different in each condition.
b) To determine noise levels of biological replicates.
c) To determine the quality of the sequence output.
d) To determine if a biological process is enriched in a list of genes induced after treatment.
e) All of the above.

8. Which of the following is NOT part of a regulatory network?

a) The A transcription factor represses gene C.
b) The A transcription factor activates gene B.
c) The A transcription factor activates its own expression.
d) The A transcription factor is degraded after treatment.
e) The A transcription factor does not interact with gene D.

9. The same environmental signal is expected to modify gene activity in_____.

a) very similar and predictable way in all individuals of the same species
b) very similar and predictable way in all individuals of closely related species
c) various ways, both similar and distinct, in individuals of even the same species
d) various ways, both similar and distinct, in individuals of closely related species

10. Which of the following is NOT true concerning control of gene expression in eukaryotic cells?

a) Transcriptional control is the most important factor
b) Transcription factors help RNA polymerase bind to a promoter
c) Transcription activators binding to enhancers can speed up transcription
d) Part of transcriptional control includes the processing of mRNA before it leaves the nucleus
e) All of the above are correct

11. Two different proteins produced by alternative splicing are most likely to:

a) Have completely different functions.
b) Have identical functions.
c) Have similar functions under different forms of regulation.
d) Have different functions under similar forms of regulation.

12. How many sequence reads are typically produced in one RNA-Seq experiment?
    a) about 10
    b) about 1000
    c) about 100,000
    d) about 10,000,000

13. What is the correct order of the steps for a typical RNA-Seq experiment?
    a) RNA extraction, cDNA synthesis, sequencing, read trimming and quality control, alignment of reads to genes, counting the number of reads for each of the genes, finding differentially expressed genes
    b) RNA extraction, sequencing, alignment of reads to genes, counting the number of reads for each of the genes, cDNA synthesis, read trimming and quality control, finding differentially expressed genes
    c) DNA extraction, RNA synthesis, sequencing, read trimming and quality control, alignment of reads to genes, counting the number of reads for each of the genes, finding differentially expressed genes
    d) DNA extraction, RNA synthesis, sequencing, alignment of reads to genes, counting the number of reads for each of the genes, finding differentially expressed genes, quality control of identified genes

14. What proportion of genes likely change their expression levels in response to environmental stress?
    a) less than 10 genes
    b) less than 0.1% of all genes
    c) 0.1 – 1 % of all genes
    d) 1 - 10% of all genes
    e) more than 10% of all genes

15. Let's say you performed RNA-Seq experiments that compared heart tissue from a strain of mice carrying a particular mutation and a wild type strain of mice. You are interested in a particular gene X and you found out that the number of reads mapped to gene X in a mutant sample is 200, while the number of reads mapped to gene X in a wild type sample is 400. Can you make a conclusion that gene X is expressed at lower level in a mutant strain?
    a) Yes, because 400 is a higher number of reads than 200
    b) Yes, because mutant strains will always have lower levels of gene expression
    c) No, there is a problem with the data because gene X should have produced no reads at all in the mutant strain due to the mutation
    d) No, because the read counts should be normalized to the length of the gene
    e) No, because the read counts should be normalized to the total number of reads generated from the samples

16. What is the fundamental difference between RNA-Seq analysis and traditional methods of investigating gene activity levels (qRT-PCR)?

a) RNA-Seq allows comparing gene activity levels for all genes at the same time, while traditional methods investigate several genes at a time.

b) RNA-Seq analysis measures RNA levels directly, while traditional methods convert RNA into cDNA first, interfering with the RNA quantities

c) RNA-Seq analysis allows for very precise quantification of RNA molecules while traditional methods, even qRT-PCR, lack the power to quantify RNA levels

d) RNA-Seq analysis is a modification of traditional methods that uses computers and technology, while traditional methods are based on "old school" pipetting

17. What IS NOT necessary to have in order to perform an RNA-Seq experiment?

a) A more or less complete genome sequence of an organism that you work on

b) Access to a genomics core facility that could make sequencing libraries and perform next generation sequencing

c) A robust protocol for isolating RNA from the tissue of the organism you work on

d) A lot of money to buy several computer software programs that are essential for data analysis

e) Two or more biological samples that could be compared

18. What is NOT true about RNA molecules that are "sequenced" during RNA-Seq experiments?

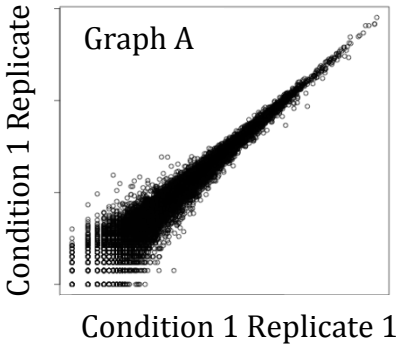
a) Most of the RNA molecules selected to be similar in size prior to the experiment

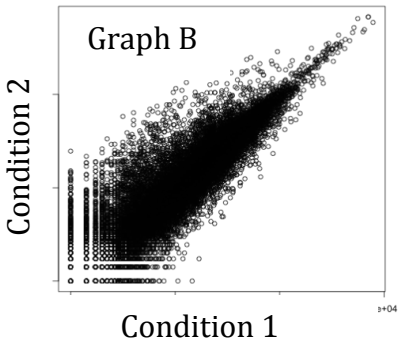
b) These RNA molecules are mostly mRNAs

c) During library construction, the RNA molecules are fragmented into 250 – 350 bp fragments

d) These RNA molecules should show no or very little degradation

19. Two graphs below show the comparison of normalized gene counts produced by an RNA-Seq experiment. The first graph (A) shows comparison of two replicates of the same condition, while the second graph (B) shows comparison of two different conditions. What can you conclude based on these graphs?





a) The data could not be analyzed further because gene counts for two replicate samples (graph A) do not all fall on a line

b) The data are of reasonably good quality because the gene counts for two different conditions (graph B) show higher variation than the gene counts for two replicates of the same condition (graph A)

c) The experiment likely was conducted incorrectly because too many genes show variation in gene expression levels between different conditions (graph B)

d) These graphs are not an appropriate tool to assess the quality of RNA-Seq data

e) I am not able to answer this question without guessing

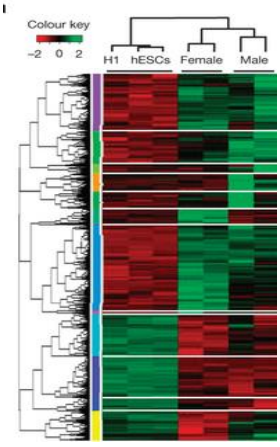
20. You are interested in identifying genes that are up-regulated in a mutant relative to a wild type organism. You conduct an RNA-Seq experiment using three replicates of RNA extracted from tissue of mutant and wild type organisms. Analysis you conduct found over 1,000 genes up-regulated in a mutant. What parameters would you likely use to narrow down the list of up-regulated genes?

a) Increase the level of statistical significance (from 1% to 0.1%, for example)
b) Increase the threshold for the ratio of expression differences (for example, from over 1.5-fold increase to over 2-fold increase in gene expression)
c) Filter out genes that are expressed at low levels in both samples
d) All of the above
e) None of the above, it is not necessary to narrow down the list of up-regulated genes, all of them are equally important

21. You conduct RNA-Seq analysis and find several hundred genes that are differentially expressed between two conditions you compared. What would NOT likely be your next step?

a) Publish a manuscript
b) Find possible functions of the differentially expressed genes and identify pathways over-represented among these genes
c) Use available resources and analyze the distribution of differentially expressed genes along the chromosomes of your species
d) Validate the expression of several differentially expressed genes using RT-PCR

22. This is a heat map constructed based on the data from RNA-Seq experiment. What is the most logical and complete interpretation of this figure?



a) The figure shows expression of seven genes (columns) across multiple samples (rows) and these seven genes fall into two groups based on their expression levels.
b) The figure shows expression of several hundred genes (rows) across seven samples (columns). Samples fall into two distinct groups that each has small differences.
c) The figure shows expression of several hundred genes (rows) across seven samples (columns). Both samples and genes are clustered into groups based on the gene's expression patterns.
d) I cannot answer this question without guessing.

Understanding Transcriptional Response to Abiotic Stress: Large Data Analysis in the Classroom

Pre-Course Survey of Student Experience with Data Analysis Approaches

Dear Students! Please rate your experience of using the following data analysis approaches. Please use a three-point rating system.

"No experience" – you have never been engaged in a project or laboratory exercise utilizing this approach of data analysis. **"Some experience"** – you participated in one or two laboratory exercises utilizing this approach of data analysis. **"A lot of experience"** –you participated in three or more laboratory exercises utilizing this approach of data analysis or used it as a part of collaborative research experience.

- Gene expression analysis, like RT-PCR, quantitative RT-PCR, RNA-Seq, or microarrays.
- Analysis of large data sets, "big data analysis," such as RNA-Seq, metagenomics analysis, genome analysis, or analysis of any other data set with thousands of data points.
- Using the software package R or other computational approaches excluding Excel for analysis of biological data.
- Amplifying and analysis of DNA fragments, like PCR and gel electrophoresis

Understanding Transcriptional Response to Abiotic Stress: Large Data Analysis in the Classroom

Rubric used for the assessment of student skills in data visualization (lab reports)

	1: Incomplete	2: Developing	3: Accomplished	4: Exemplary
Experimental question	No experimental question is stated	Question is too general or too specific, no justification by the introduction	Question is clearly stated; however, it is not justified by the introduction	Question is clearly stated and justified by the introduction
Graphs	No graphs are present or the graphs do not illustrate RNA-Seq data	Chosen graph types are not appropriate for the data or do not show data relevant to the stated question; graphs have several mistakes	Chosen graph types illustrate the data relevant to the experimental question; graphs have minor mistakes	Chosen graph types illustrate the data relevant to the experimental question; graphs show correct data
Graph labels	All graph labels are missing or not clear	Most of the necessary graph labels are missing or not clear	Some of the necessary graph labels are missing or not clear	Graphs contain all appropriate labels (axes, figure legend, etc…)
Figure legends	Figure legends are not present. Figures do not have titles.	Figure legends are too general or too detailed. Important information is missing or unclear. Many figures do not have clear titles.	Figure legends are precise and concise and describe the data and the graphs. Some important information is missing or unclear. Some figures do not have clear titles.	Figure legends are precise and concise and clearly describe the data and the graphs. Figures have titles.
Data interpretation	Very incomplete or incorrect interpretation of trends and comparison of data indicating a lack of understanding of results.	Some of the results have been correctly interpreted and discussed; partial but incomplete understanding of results is still evident. Specific connections to the experimental question are missing.	Most of important trends and data comparisons presented on the graphs are interpreted correctly and discussed in relation to the experimental question. Good understanding of results is conveyed.	All important trends and data comparisons presented on the graphs are interpreted correctly and discussed in relation to the experimental question; good understanding of results is conveyed.

Supplemental Table 1. List of Experimental Questions chosen by the Students and the Approaches Students Used to Visualize Relevant Data

Experimental Question	Data Visualization Approaches
Expression of how many genes is affected by cold?	Tables, Venn diagrams, lists of DE genes
What biochemical pathways are activated in response to stress?	Bar graphs of Gene Ontology annotations
Do different abiotic stress conditions elicit similar or different responses in gene expression?	Bar graphs, heat maps, Venn diagrams, scatter plots
How conserved is the transcriptional response to stress between different maize inbreds?	Heat maps, Venn diagrams, scatter plots
How conserved is the transcriptional response to stress between individual plants?	Heat maps, scatter plots
What genes are activated in response to any stress?	Heat maps, combining lists of DE genes, Venn diagrams
What types of genes respond to abiotic stress in maize?	Bar graphs of Gene Ontology annotations

Supplemental Table 2. Number of Genes Differentially Expressed in Response to Abiotic Stress

Number of genes	Cold[a]		Heat[a]	
	B73	Mo17	B73	Mo17
FGS genes (37,537)				
Up-regulated genes[b]	2,984 (7.9%)	3,636 (9.7%)	3,648 (9.7%)	4,287(11.4%)
Down-regulated genes[c]	2,572 (6.9%)	3,259 (8.7%)	3,839 (10.2%)	2,214 (5.9%)

[a]Analysis of cold and heat stress conditions is based on three biological replicates for all genotypes

[b]Up-regulated genes were identified as genes with expression level at least 2 fold higher under stress condition compared to control, expressed at least at RPM level of 1 under a stress condition, and with a statistical significance level of at least 0.05

[c]Down-regulated genes were identified as genes with expression level at least 2 fold lower under stress condition compared to control, expressed at least at RPM level of 1 under control conditions, and with a statistical significance level of at least 0.05

Supplemental Figure 1. Student analysis of the quality of sequence runs for the data set on plant response to abiotic stress. (A) An example of a Per Base Sequence Quality report generated by FastQC application in Green Line of the DNA Subway. For all 50 nucleotides of the reads, majority of the sequences have high quality scores. (B) An example of a Per Sequence Quality Scores report generated by FastQC application in Green Line of the DNA Subway. Majority sequences from this sample have high quality scores.