Bio 311 • Group HW #1 • 30 March 2017

Group 1: Karen Li, Sam Yin, Daniel Zhu, Lauren Shum

Describe your biological question and project scope. For this clustering assignment, did you choose to work with the entire data set, or a subset of the data? How did you make this choice?

- Biological question posed by researchers: The original researchers were trying to determine the relationship between temporal changes in nucleosomal position, transcription factor binding, and gene expression following the introduction of a stressor (hydrogen peroxide).
- Project scope: Our data consists of a ChIp data with the expression levels of about 11,000 genes over time in wild type and in the Msn2/4 double deletion yeast. From this data, we can cluster those genes into subgroups based on the similarity of expression patterns with and without functioning copies of Msn2/4, which would allow us to examine how various genes respond to the Msn2p/4p-regulated stress response. The provided data set, however, does not include information about nucleosomes, which was used by the original researchers to show that regulatory changes often precedes nucleosome modification rather than the other way around. Therefore, we chose to focus on a different biological question.
- Our biological question: We want to determine whether and how the gene expression for wild type strains is different from that for yeast strains with Msn2 and Msn4 deletions when both are exposed to hydrogen-peroxide.
- Choice of data set: Because the original reference aimed to explore functional differences in handling stress, we chose to cluster with the subset of coding genes only, since genes that produce proteins are more likely to be relevant to the stress response process. (However, if we were to observe nucleosome occupancy, we would observe gene expression levels for both coding and non-coding genes as nucleosomes are present on both.) We also removed any genes with missing data at any time points (i.e., N/A instead of numeric value); the purpose of this was to remove the possibility of nonexistent correlations between any two genes.

Describe your clustering methods.

- Methods & distance metrics: We attempted both hierarchical and K-medoid clustering. The hierarchical clustering was done with a cut-height of 1.8 using pairwise complete complete linkage using the hclust algorithm. The cut-height of 1.8 was chosen largely because it produced a reasonable sized clusters after averaging repeated data points. On average our clusters contained approximately 500-600 genes, with the smallest cluster containing 227 genes, and the largest cluster containing 1489 genes. Our clustering resulted in 9 groups containing a significant number of genes. Thus, we will use this information to determine relationships within and between clusters utilizing further methods.
- Contrasting different k values: For the k-means clustering, we generated two different clusterings for comparison, and for our determination of the k value. In figure 3, we initially set k=11 for comparison. This clustering showed that the smaller clusters were distinctly grouped, while other clusters such as cluster 3 seemed to express similar clustering characteristics as cluster 4, and contained more in-cluster variability. This

contrasted with our second clustering where k=9. In figure 4, we set k to be 9 so that the results would be comparable to the results found by hierarchical clustering. As shown in our figure, the uniformity and magnitude of the clusters on the diagonal of our k-medoids heat map shows a relatively stronger within-cluster similarity. We also observe that relationships between clusters is relatively easier to define in our 9-clustering as opposed to a higher clustering.

- "Best" clustering methodology: Between the two, the k-medoid clustering appeared to produce clusters that were more similar in terms of temporal patterns. Since we are mostly interested in expression patterns that only emerge over time, we decided that this would be the best clustering for our purposes.

What is the distribution of sizes of clusters in your analysis (few clusters with many genes or many clusters with few genes)?

- Many clusters with few genes. Our clustering analysis found about 20 clusters that split off from each other very early on. While those clusters can be further subdivided, it is very difficult to meaningfully group them together into a smaller number of clusters.

If your focal paper(s) described cluster analyses of the expression data, how do their results compare to yours?

- Focal paper results: The researchers used k-means clustering with k = 4 to classify genes according to the timepoint at which nucleosomes were gained or lost along with the relative nucleosome occupancies at different timepoints. Further clustering methods involved classifying nucleosomes based upon its gain/loss dynamics as a function of the time they were exposed to H_2O_2.
- Our results: Because our dataset did not provide any data referring to nucleosome occupancy, our biological question and, therefore, our results, describe gene expression differences between wild type yeast and yeast strains with Msn2/4 double-deletions, and are not very comparable to the results of the focal paper. We found a set of 9 clusters that have consistent within-cluster temporal patterns with and without the Msn2/4 deletions.

Bio 311 • Group HW #1 • 30 March 2017

Group 1: Karen Li, Sam Yin, Daniel Zhu, Lauren Shum

Generate an overview figure for your clustering analysis that you can use for your poster. In your figure, include a heatmap and dendrogram (if you used hierarchical clustering), or other graphical representation of your clustering analysis.

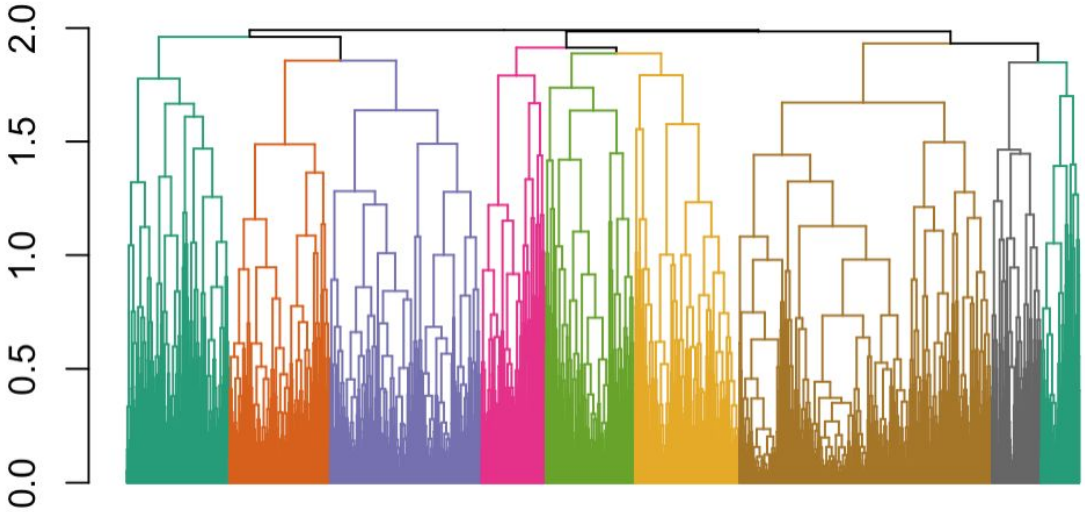


Figure 1: Hierarchical clustering of data set, cut height = 1.8.

Table 1: Cluster sizes for hierarchical clustering, cut height = 1.8.

Cluster #	1	2	3	4	5	6	7	8	9
Cluster size	604	594	893	378	526	615	1489	287	227

Table 2: Cluster sizes for k-medoids clustering, k = 9.

Cluster #	1	2	3	4	5	6	7	8	9
Cluster size	455	803	764	370	877	552	571	676	545

The average size of clusters for both clustering algorithms was 624 genes. The standard deviation for the hierarchical clustering was 381, whereas it was 168 for k-medoids clustering. This indicates that the cluster sizes were more uniform for k-medoids clustering.

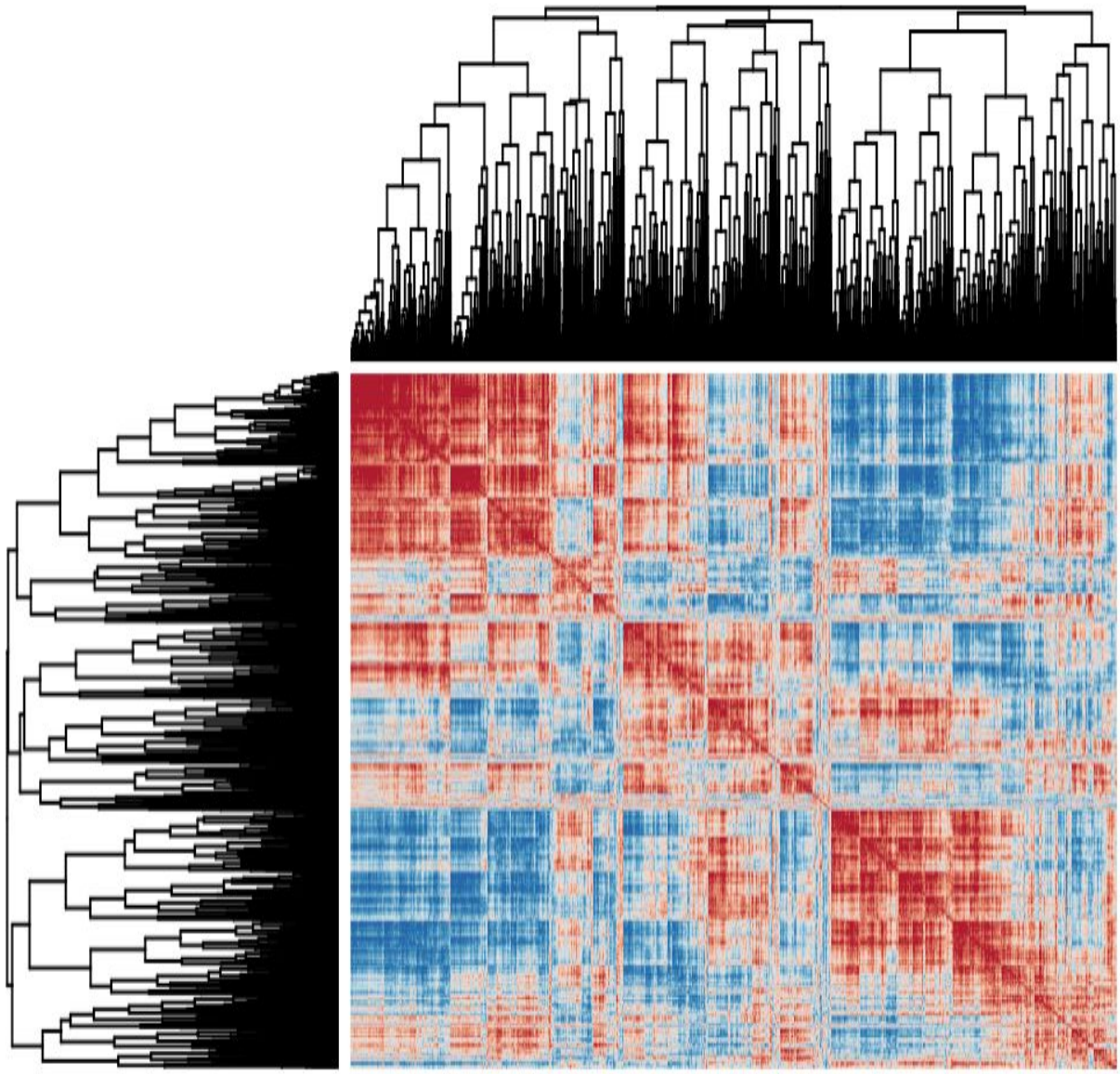


Figure 2: Gene correlation matrix heatmap sorted by hierarchical clustering, cut height = 1.8.

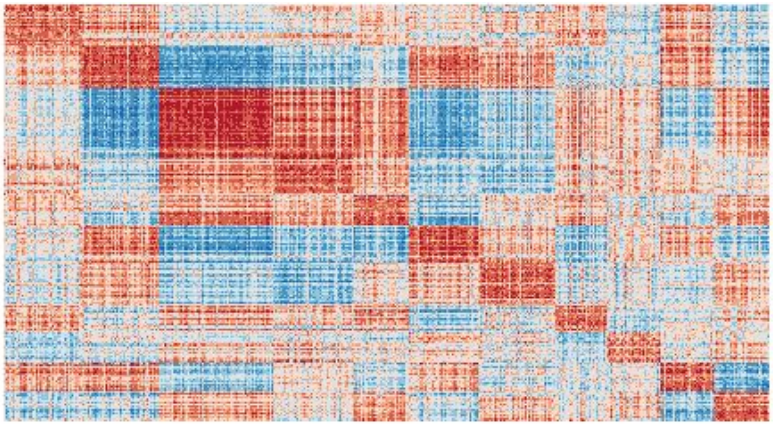


Figure 3: Gene correlation matrix heatmap sorted by k-medoids clustering, k = 11.

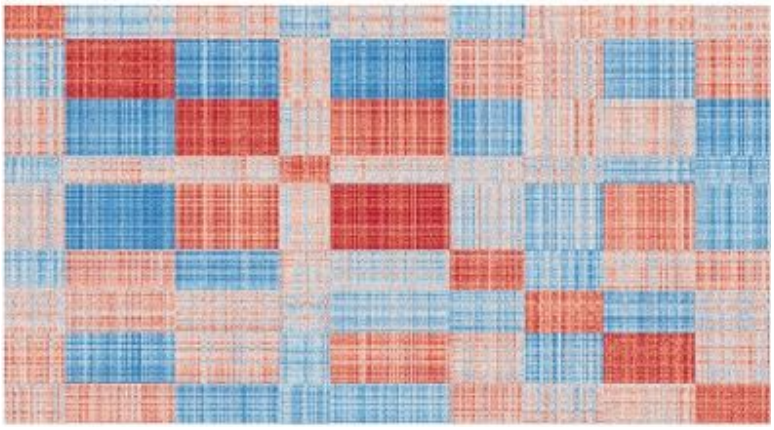


Figure 4: K-medoid clustering with k = 9 (averaging repeated experiments at 30 minutes)

Bio 311 • Group HW #1 • 30 March 2017

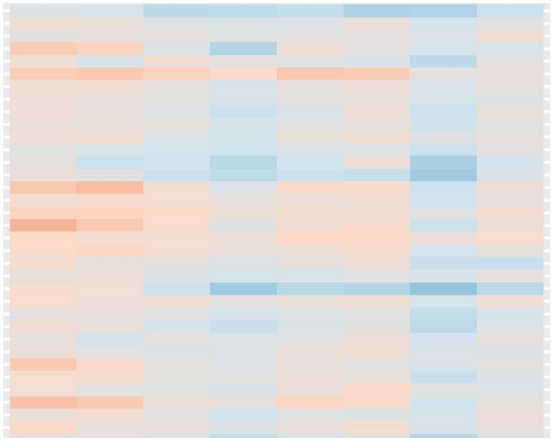
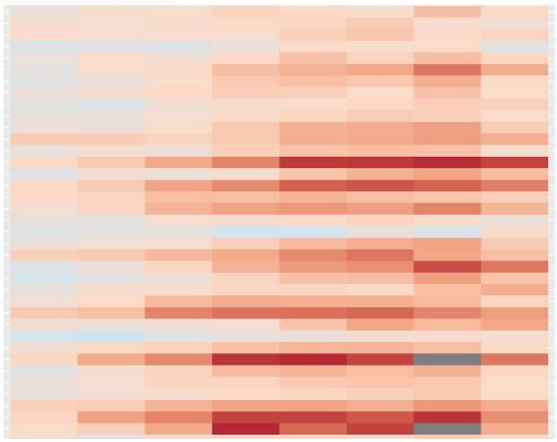
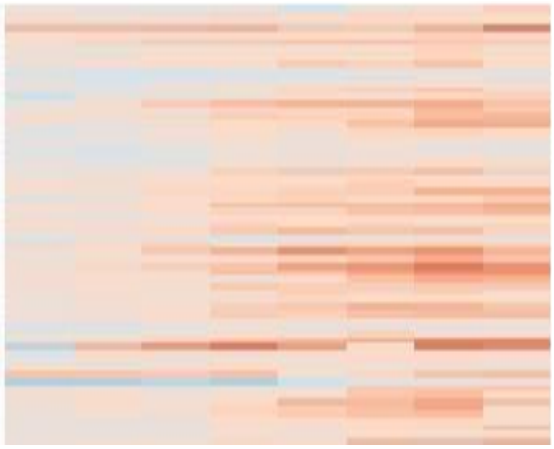
Group 1: Karen Li, Sam Yin, Daniel Zhu, Lauren Shum

Comparison between wild type and Msn2/4ΔΔ gene expression dynamics for each k-medoid cluster:

Expression

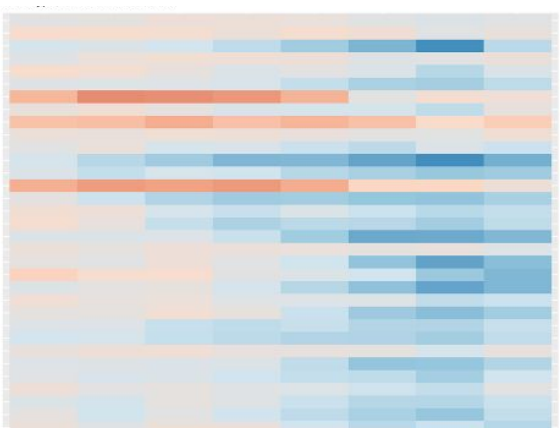
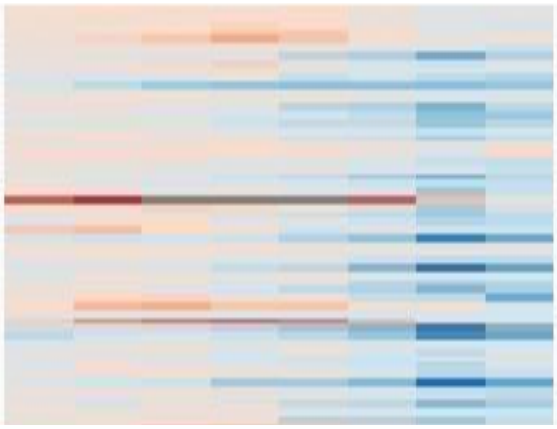
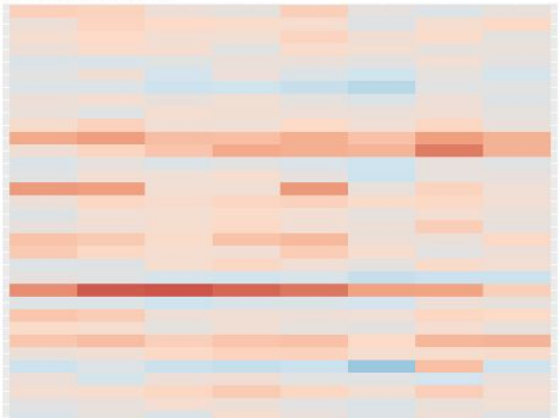
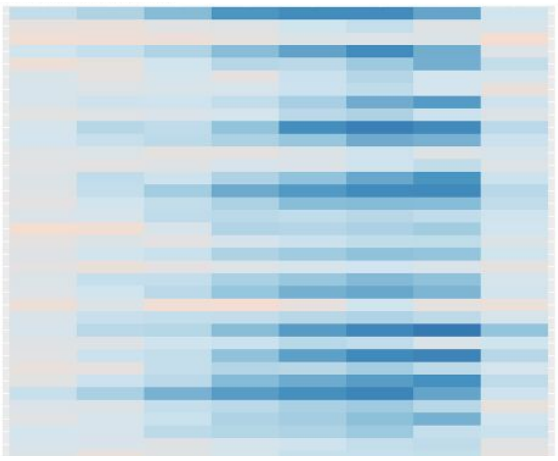
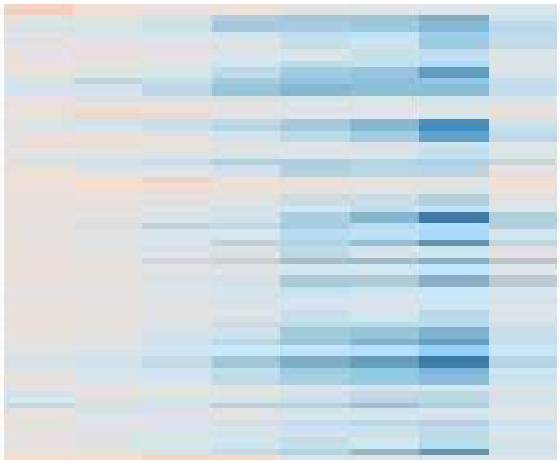
4 2 0 -2 -4

(gray means expression >4)

	Wild Type	**Msn2/4 ΔΔ**
1		
2		

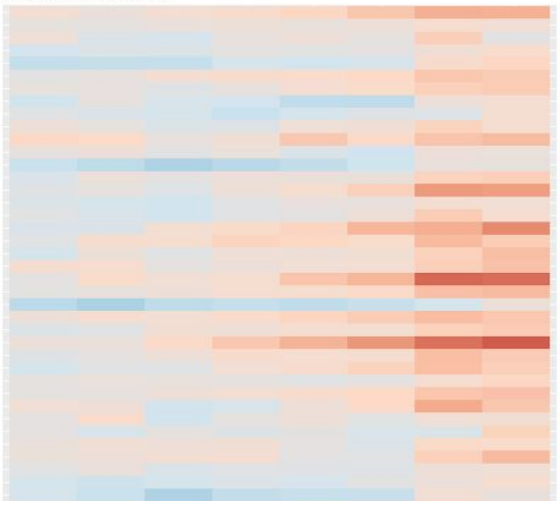
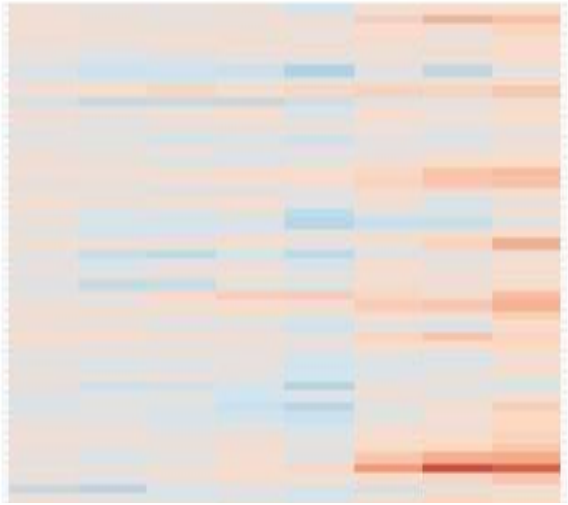
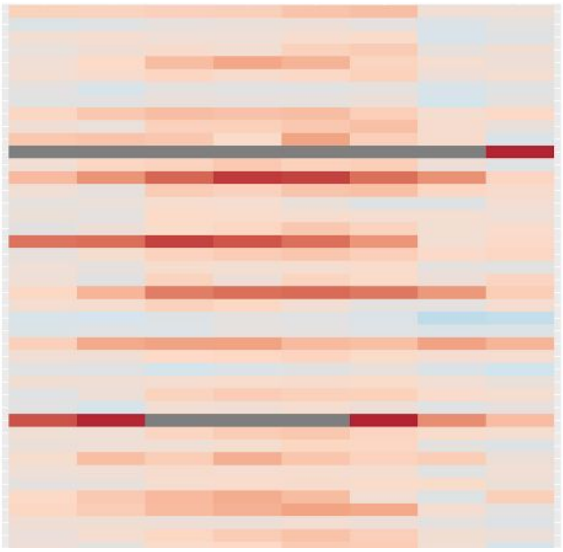
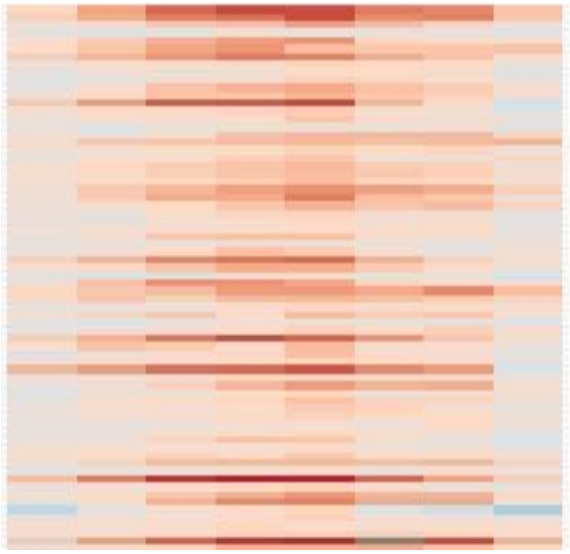
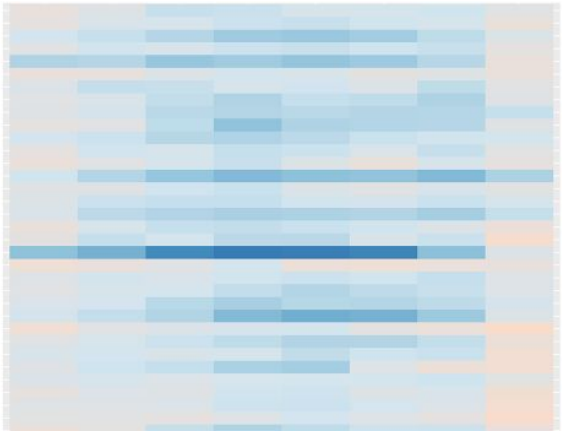
Bio 311 • Group HW #1 • 30 March 2017

Group 1: Karen Li, Sam Yin, Daniel Zhu, Lauren Shum

3		
4		
5		

Bio 311 • Group HW #1 • 30 March 2017

Group 1: Karen Li, Sam Yin, Daniel Zhu, Lauren Shum

6		
7		
8		

Bio 311 • Group HW #1 • 30 March 2017

Group 1: Karen Li, Sam Yin, Daniel Zhu, Lauren Shum

9