The first step in verifying RosR’s function in regulating halobacteria’s adaptability to highly oxidative environment, we carried out clustering analysis on the oxidative stress data set, comparing the gene expression levels of RosR knockout cells with those of the parent strain. We hypothesized that if we were able to find clusters within the RosR data set that show a drop in gene expression compared to that of the parent strain set, we would validate the fact that RosR is responsible for regulating oxidative stress response in Halobacterium.

We decided to start out with k-means clustering as it is one of the simplest and most robust algorithm for unsupervised learning. To determine the appropriate number of k-clusters, we looked for an “elbow” in the sum of squared error scree plot, which suggest stability and convergence of the clusters over many iterations. This analysis suggested that 10 was a suitable number of clusters for the given dataset. Then, we made a heatmap of the parent strain and the RosR deletion strain separately, reordering the genes by the time they reached peak expression. The vertical columns are the separate time-series points of the parent and RosR deletion strains and the horizontal rows are individual transcription factors provided in the dataset. We understand that the figure is bit small, but we wanted to highlight the difference between the two by aligning them side by side.

(A) genes requiring VNG0258H for repression regardless of condition, (D) genes requiring VNG0258H for activation in the presence of H2O2, (F) genes requiring VNG0258H for impulse-like dynamic induction, (G) genes requiring VNG0258H for impulse-like dynamic repression, (H) genes induced in response to H2O2 but independent of VNG0258H, and (J) genes repressed in response to H2O2 but independent of VNG0258H. We wanted to see if the genes we found using the clustering analysis also were found in the expression profiles noted in Sharma et al. In this process, we were trying to validate whether or not our clustering results were due to the fact that we were missing data points.