# Verifying the Role of RosR Transcription Factor in the Oxidative Stress Response of *Halobacterium Salinarum* NRC-1

Yitaek Hwang, Tosin Omofoye, Steven Pierre, Zhihao Zhu

## Abstract

Previous work by Sharma et al. and Tonner et al. have shown that VNG0258H or RosR is a transcription factor that regulates gene expression in *Halobacterium Salinarum* NRC-1 under extreme oxidative stress. The goal of this project was to verify the results of the aforementioned papers and delve deeper into identifying specific genes that RosR either activates or represses in response to oxidative stress. This allows to map the genes whose expression are directly affected by RosR. In comparing the gene expression ratio in the *Δura3* parent and *ΔVNG0258H* mutant strains before and after exposing each subject to $H_2O_2$, we validated six expression profiled noted in Sharma et al. Then, we used k-means clustering to find a list of genes regulated by RosR using the same gene expression profile. This list, however, differed significantly from the list provided in Sharma et al., leading us to question our stringent criteria for ChIP-chip analysis. Next, we created a transcription factor network to find the relationship between the duration of $H_2O_2$ exposure to the genes that RosR regulates. The specific functions of these genes were then analyzed using the arCOG gene ontology analysis. Lastly, we compared our results with those of Tonner et al. to validate our analysis. In conclusion, the analysis showed that there are very few genes that have very high correlation to RosR binding at every stage of response to extreme peroxide levels, but many genes do have high correlation to RosR binding immediately after exposure to extreme levels of oxidation.

## Clustering Analysis

We used k-means clustering to start exploring the data as it is a simple and robust clustering algorithm for big datasets. To determine the appropriate number of k-clusters, we looked for an “elbow” in the sum of squared error (SSE) scree plot. The location of the “elbow” suggested 10 as a suitable number of clusters for the k-means analysis of parent strain dataset. A dendrogram was created based on these k-means clusters.

Heatmap based on peak expression levels of each gene was created for the parent strain as well as RosR deletion strands.The gene expression data values are in $\log_{10}$ values. We exponentiated this dataset to get the nominal values and then found the differences between the maximum gene expression in the parent and RosR deletion strains. Genes with differences outside 2 standard deviations of the mean difference have significant change in gene expression due to RosR.

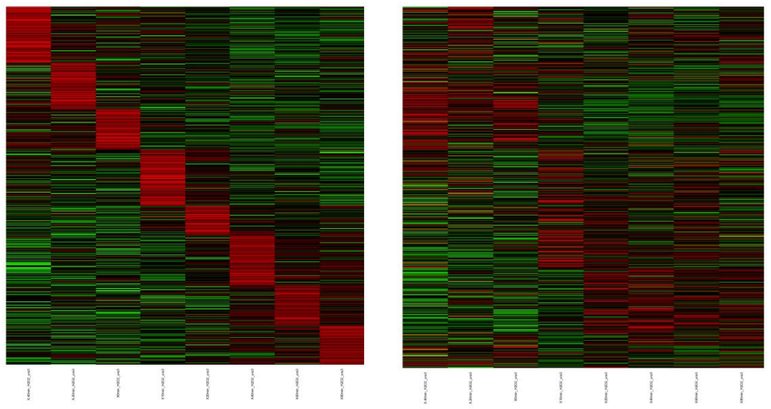


Figure 1: Maximum Gene Expression for Each Gene over the 8 Time Points

We computed the maximum gene expression over time for each gene in the data set and reordered based on peak expression profiles. The left heatmap in Figure 1 shows the peak gene expression (y-axis: genes, x-axis: time points) for the parent strain, whereas the right heatmap in Figure 1 shows the peak gene expression for the RosR deletion strain experiments.

Using all the differences between maximum expression of genes in the parent strain and RosR deletion strain experiments, we found 142 genes that exhibit significant differences in gene expression. The identified genes were then compared against the list in Sharma et al. paper first and then analyzed further using transcription factor network analysis and gene ontology.

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| VNG0218G | VNG0339H | VNG0349G | VNG0357H | VNG0462C | VNG0506H | VNG0509H | VNG0878G |
| 176 | 261 | 268 | 273 | 352 | 388 | 391 | 670 |
| VNG0931G | VNG0932C | VNG0978H | VNG1087C | VNG1347C | VNG1366H | VNG1497C | VNG1547C |
| 710 | 711 | 746 | 833 | 1029 | 1043 | 1143 | 1179 |
| VNG1583C | VNG1674H | VNG1748C | VNG1872C | VNG1956H | VNG1997G | VNG2094G | VNG2164C |
| 1206 | 1270 | 1323 | 1418 | 1483 | 1514 | 1585 | 1646 |
| VNG2338G | VNG2440H | VNG2471G | VNG2472G | VNG2593H | VNG2594C | VNG5039G | VNG5059C |
| 1788 | 1866 | 1890 | 1891 | 1979 | 1980 | 2074 | 2086 |
| VNG5083H | VNG5090C | VNG5091C | VNG5146H | VNG5157H | VNG5160H | VNG5165H | VNG5167H |
| 2104 | 2109 | 2110 | 2143 | 2151 | 2152 | 2155 | 2157 |
| VNG6223C | VNG6284H | VNG6312G | VNG6323H | VNG6329H | VNG6330H | VNG6335H | VNG6339H |
| 2250 | 2291 | 2312 | 2322 | 2326 | 2327 | 2330 | 2332 |
| VNG6353H | VNG6393H | VNG6411H | VNG6416H | VNG6418H | VNG6424H | VNG6432H | |
| 2341 | 2369 | 2381 | 2384 | 2385 | 2388 | 2393 | |

## Replicating Sharma et al. Results

Gene expression in response to $H_2O_2$ exposure in the *Δura3* parent and *ΔVNG0258H* mutant strains are shown below. Like Figure 5 in Sharma et al., each line in the graph detail the $\log_2$ gene expression ratio for time points before and after the $H_2O_2$ exposure noted by the dotted line:

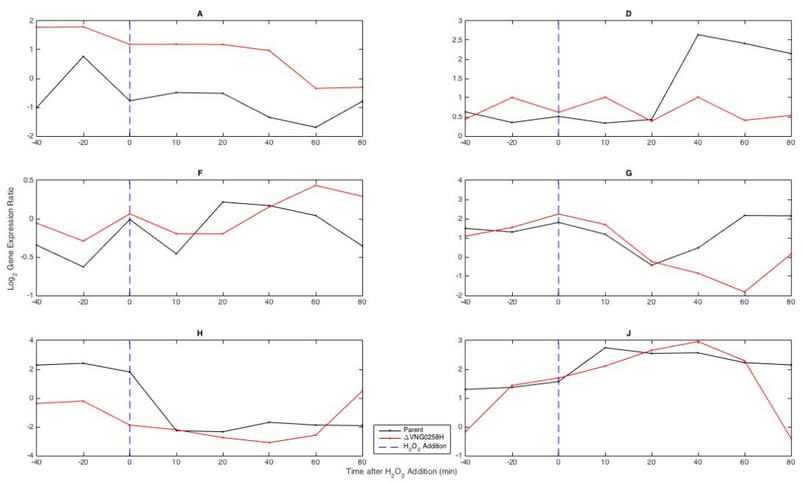



Figure 2: Gene Expression Profiles from Sharma et al.

(A) genes requiring VNG0258H for repression regardless of condition, (D) genes requiring VNG0258H for activation in the presence of $H_2O_2$, (F) genes requiring VNG0258H for impulse-like dynamic induction, (G) genes requiring VNG0258H for impulse-like dynamic repression, (H) genes induced in response to $H_2O_2$ but independent of VNG0258H, and (J) genes repressed in response to $H_2O_2$ 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.

## TF Network Analysis

The ChIP-chip data provided detailed the interactions of a list of genes in response to the addition of $H_2O_2$ at various time points. We used a threshold of 0.001 on the p-value to find the genes with high probability of significant interaction with RosR. We chose this very small cutoff value as we wanted to focus on smaller group of genes that have very high relation to RosR and study their functions and interactions to produce the response to reactive oxygen stimulus. Next an incidence matrix was made using the data at the time points following addition of $H_2O_2$. The transcription factor network generated below shows the significant interaction between RosR and individual genes at specific time points after the exposure to oxidative stress.

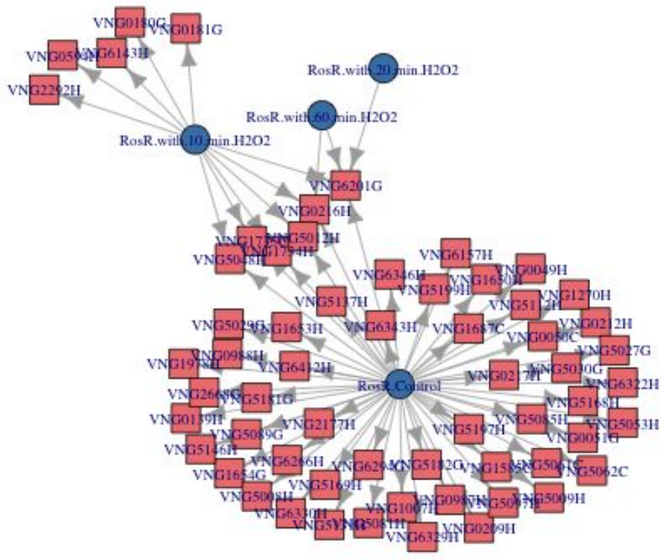
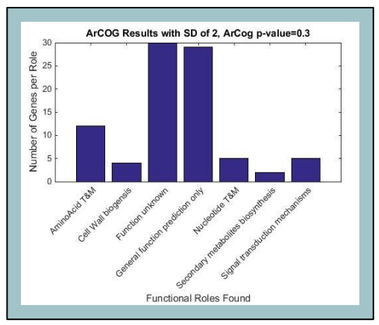


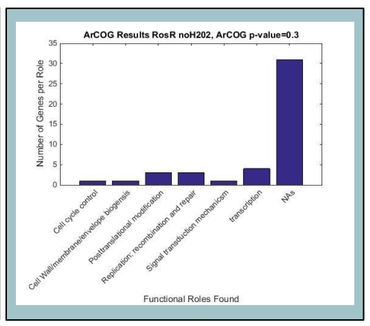

Figure 3: Gene Network for Gene Regulated by RosR over Time

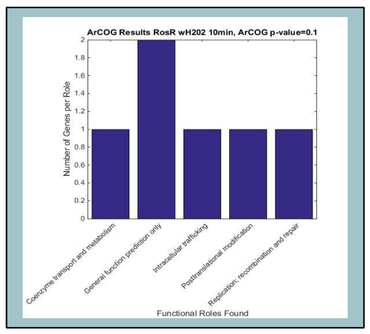
Only one gene, VNG6201G, was regulated by RosR at every time point. This gene, also known as hsp5, is a heat shock protease protein and also related to peptidase activity. At a small p-value cutoff, it was hard to determine the relationship between RosR and the duration of oxidative stress levels.

## Gene Ontology Results

Gene ontology analysis was performed on our gene expression data and transcription factor network data using arCOG (archaeal Clusters of Orthologous genes, Wolf et al., 2012). Based on our max expression difference histogram data, arCOG was performed on genes outside of a standard deviation of 2. The results are presented in the top-left plot of the figure below. For the TF network data, arCOG was performed using p-value of 0.2 on the sets of genes that strongly interacted with RosR before $H_2O_2$ was added, after 10 minutes and then after 60 minutes. arCOG could not be performed on the 20 minute set since only one gene was present.




ArCOG analysis on genes that showed a significant change in max expression levels between the parent strain and the RosR deletion strain. (Top left) ArCOG analysis on genes that have a strong interaction with RosR before $H_2O_2$ addition (Top right). ArCOG analysis on genes that have a strong interaction with RosR 10 minutes after $H_2O_2$ addition (Left)

One resultant functional role that we observed was “replication: recombination and repair”. This role showed up as a category twice in our arCOG plots; it is relevant because we can expect RosR to play a role in repairing the damage caused to DNA by oxidative stress. We also found that there were a lot of genes with unknown or only predicted functions. We expect this to change as we find out more about gene with currently unknown functions.

## Conclusions

In conclusion, we found that RosR influenced the expression of 142 genes through the means of our k-means clustering analysis. In turn, we also observed during our transcription factor network analysis that only one of those genes was affected directly by RosR over the whole exposure time. Although our results differ significantly with Sharma et al. and Tonner et al., we see through our replication of the gene expression data that RosR regulates some genes under $H_2O_2$ stress. The reason for this difference may be from our lack of binding sequence motifs or from our stringent p-value cutoff. Also, in comparison to the Tonner et al. paper, we do see that RosR activates and represses certain genes, throughout the entire exposure time period, while most genes are regulated immediately after the addition of $H_2O_2$. The arCOG analysis did not reveal significant function of any of these genes, but we are confident that in the future, we will know more about currently unknown functions.

## References

[1] Sharma, K., Gillum, N., Boyd, J. L., & Schmid, A. (2012). “The RosR transcription factor is required for gene expression dynamics in response to extreme oxidative stress in a hypersaline-adapted archaeon.” *BMC genomics*, *13*(1), 1.

[2] Tonner, P., Pittman, A., Gulli, J., Sharma, K., & Schmid., A. (2015). “A Regulatory Hierarchy Controls the Dynamic Transcriptional Response to Extreme Oxidative Stress in Archaea.” PLos Genet. 2015 Jan. 11(1): e1004912.

[3] Wolf, Y., Makarova, K., Yuin, N., & Koonin, E. (2012). “Updated clusters of orthologous genes for Archaea: a complex ancestor of the Archaea and the byways of horizontal gene transfer.” *Biology Direct,* 7:46.