Raehoon Jeong
Biophysics 205 Final Proposal

Dynamic changes in transcription during heat shock response in human and mouse

Introduction

Cells can go through numerous types of stress throughout their lifetime, and they harbor multiple mechanisms to counter their effects the best they can. One of them is heat shock. Even though various organisms have adapted to survive in a wide range of temperatures, a mere few degrees rise in temperature above the optimum can inflict heat stress to a cell. Some of the detrimental effect from heat stress include misfolding and aggregation of proteins, defects in cytoskeleton, aberrant splicing, and disruption of ion homeostasis due to changes in membrane permeability (Richter et al., 2010) (Fig. 1). These radical alterations are harmful to the cells, so they initiate a rapid transcriptional program called the heat shock response in order to minimize the damage. The fact that such slight difference in temperature launches a significant change in gene expression makes it an interesting phenomenon to study to further understand gene regulation. Furthermore, applying heat stress to cells is as easy as incubating them in a higher temperature setting.

Since correctly refolding damaged proteins conserves more energy than transcribing and synthesizing the protein from scratch, a large component of the heat shock response involves activating molecular chaperones to guide proper refolding (Gidalevitz et al. 2011). This family of proteins with multiple members are aptly named heat shock proteins. These proteins earned this name because they are encoded by genes that are upregulated under heat stress. The first observation of such gene activation in response to heat stress was in *Drosophila*, (Jamrich et al., 1977) and afterwards, there came reports of other organisms from prokaryotes to eukaryotes also exhibiting a similar phenomenon (Kelley & Schlesinger, 1978; Lemaux et al. 1978; McAlister & Finkelstein, 1978). Since the heat shock response is such a universal behavior of a living cell, studying the underlying regulatory mechanism reveals basic principles.

Upregulation of heat shock protein genes occur in a matter of minutes, so it was obvious that a transcription factor would play a key role (Vihervaara et al. 2018). In eukaryotes, this key transcription activator is heat shock factor 1 (HSF1). In cells under normal condition, HSF1 is inhibited by heat shock proteins bound to it in the cytoplasm. However, when the cell experiences heat stress in the form of protein misfolding, the heat shock proteins dissociate from HSF1 to perform its chaperone function. Then, HSF1 is trimerized and transported to the nucleus to activate heat shock response genes (Vihervaara & Sistonen, 2014). Mahat et al. (2016) investigated the heat shock response transcriptional program with PRO-seq that captures nascent RNA transcripts. They confirmed that when HSF1 is knocked out in mouse embryonic fibroblasts (MEF), there was depletion of newly synthesized transcripts encoding

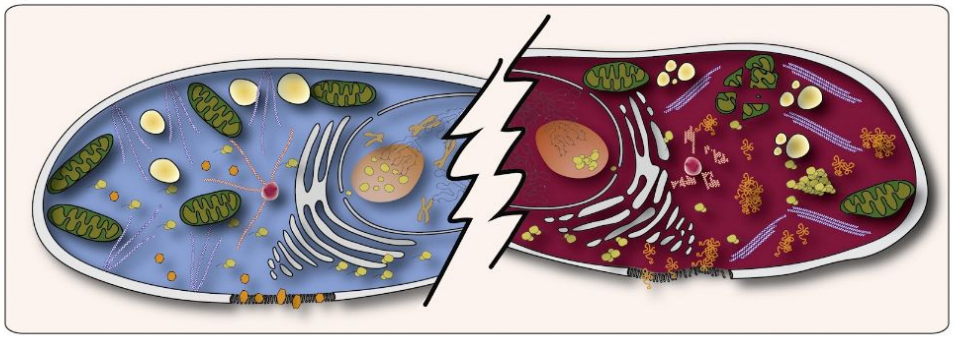

Figure 1. Damaging effects of heat stress to a eukaryotic cell. (Richter et al., 2010)
Compared to the cell in a normal condition (left), the cell undergoing heat stress exhibits various damages (right). Actin filaments (blue) and microtubules (red) rearrange and aggregate. Organelles like the endoplasmic reticulum (white) and mitochondria (green) lose integrity. And misfolded proteins form aggregates (orange).

heat shock proteins. However, there were genes that were still upregulated in absence of HSF1. These genes showed functional enrichment related to the cytoskeleton. Actually, it is consistent with previous findings that cytoskeleton proteins are integral for the response to heat stress (Baird et al, 2014). These genes seem to be regulated by another transcription factor, serum response factor (SRF), deduced from strong enrichment of its DNA motif and ChIP-seq peaks. However, instead of continuous induction, they are only transiently activated, then downregulated, reflected in nascent transcript levels as well as SRF binding. Motivated by this complexity, we hypothesize that the genomic landscape surrounding each gene also plays a critical role in the specificity of regulation, which will be dealt with in Aim 1.

Although the heat shock response is a universal response across life, there are subtle differences in the actual mechanism of gene regulation. For example, flies and human have different proteins contributing to promoter opening of heat shock activated genes (Vihervaara et al. 2018). Hence, even more closely related species like mouse and human could have minor variations in the regulatory elements. In Aim 2, PRO-seq data from varying duration of heat shock on MEF from Mahat et al. (2016) and similarly run PRO-seq experiment for K562 cells − a human leukemia cell line − will be leveraged. For MEF, a *Hsf1-/-* line will be included, and for K562, HSF1 knockdown with RNAi. We will compare the regulatory trend during a heat shock response between the two species on a gene-by-gene basis. The power of this comparison comes from the fact that human and mouse genome demonstrate high conservation of protein-coding genes but low conservation of non-coding regions. Furthermore, the temporal aspect of the data adds more resolution as demonstrated in Dukler et al. (2016) and Mahat et al. (2016).

Not only is the quick activation of molecular chaperones crucial, but the rapid recovery from a heat shock response is also equally significant. This is because the response involves downregulation of core genes related to the cell cycle, translation, and many other key biological processes (Mahat et al., 2016). Aim 3 is to utilize time-course PRO-seq experiments data from both cell lines recovering from heat stress in order to look for clusters of genes with similar transcription profiles. This approach will reveal genes that are regulated in a HSF1-dependent manner and those that have different expression patterns in the two species.

Aim 1. Analysis of the genomic landscape around genes activated during a heat shock response.

In order to narrow the candidate transcription factors regulating various clusters of similarly expressed genes during heat stress, Mahat et al. (2016) looked for enriched motifs in upregulated genes. Although this approach can be sensitive enough to identify the responsible transcription factor, it does not consider whether genes that are not upregulated also has those motifs in the promoter regions. Thus, the first aim is motivated by the possible role of the complex genomic landscape, including histone modifications, DNA accessibility, and topologically associating domains, as well as the identity of expressed TFs, in the specificity of gene activation on top of the mere presence of transcription factor binding sites.

The first part of this aim is to study histone modifications and DNA accessibility at transcription start sites of genes across various mouse cells. Data from published papers will be utilized in the analysis (Table 1). Upon heat shock, genes can be classified into 1) those undergoing rapid HSF1-dependent upregulation, 2) those undergoing rapid upregulation in an HSF1-independent manner, 3) those undergoing late upregulation, and 4) those being downregulated (Fig. 2). Using the classification of genes in Mahat et al. (2016) based on temporal expression profile throughout heat shock response, trends in histone modifications and DNA-accessibility can be studied. A simple approach would be to average the signal over each class and comparing. If there is high variation within each class, clustering the genes by similar epigenetic profiles could provide more granular information. If there is a consistent trend in some of the gene

classes, we can also test if different cell types show similar genomic landscape for these genes. Since the heat shock response is a transcriptional program that any type of cell should be able to launch, one would expect that genes that are differentially regulated during the response to be wired similarly. On ENCODE, dataset of ATAC-seq and ChIP-seq targeting various histone modifications for a range of mouse cell types are present. Consistent epigenetic marks indicate that they may have a role in their regulation. Otherwise, they are likely not relevant, and it could be that TFs, including HSF1, may act as pioneer factors.

The next part is to probe whether rapidly activated genes include TFs. Mahat et al. (2016) demonstrated that there are genes activated by HSF1 are mostly heat shock proteins and that SRF likely plays a role in transient activation. Notably, there was a class of genes that were activated later in the heat shock response, which might indicate that they are regulated by TFs that are activated in the first wave. Thus, by comparing steady state RNA-seq data of MEF from ENCODE with PRO-seq data at 2.5 and 12 minutes post-heat shock from Mahat et al. (2016), one can derive a list of TFs that are lowly expressed in steady state but is upregulated rapidly. These might narrow down candidate TFs that have a role in late gene activation.

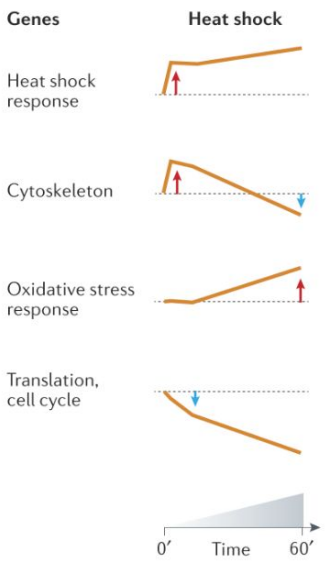

Figure 2. Temporal expression patterns during heat stress. (Vihervaara et al., 2018) PRO-seq experiments on samples from different time points during a heat shock response reveal genes of various expression patterns. The left column denotes functions that are enriched in each group.

The remaining objective is to further test the hypothesis that interactions between enhancers and promoters. Many of the genes upregulated upon heat stress in a HSF1-dependent manner did not have HSF1 binding in the promoters detected by ChIP-seq experiments. However, there were several intergenic regions with peaks corresponding to HSF1 binding that had divergent transcriptions occurring (Mahat et al., 2016), which poses the question whether they capture enhancer activities. Therefore, by incorporating Hi-C data for mouse embryonic fibroblasts (Battulin et al., 2015), whether there are putative active enhancers marked by HSF1 binding are in the same topologically associating domain as genes activated in a HSF1 dependent manner can be tested. If there are numerous examples, then we can attempt to match the enhancer to a target gene during a heat shock response.

Cell line	Data type	Type of information	Source
Mouse Embryonic Fibroblast	PRO-seq	Nascent transcription	Mahat et al., 2016
	RNA-seq	Transcription	Yue et al., 2014
	ATAC-seq	DNA accessibility	Maza et al., 2015
	ChIP-seq	Histone modifications (H3K4me1, H3K4me3 H3K27ac)	Yue et al., 2014
	ChIP-seq	HSF1	Mahat et al., 2016
	Hi-C	Chromosome conformation	Battulin et al., 2015

Table 1. Partial list of datasets that will be leveraged in the analysis

Aim 2. Comparing the heat shock response in mouse and human cells.

Since the heat shock response is a phenomenon universal to living cells, it is noteworthy to understand the similarities and the differences between species. Human and mouse have similar gene sets and comparable genome sizes at around 3 billion bases. Despite high conservation of protein-coding gene sequences, only about 3-8% of the noncoding region is estimated to undergo purifying selection (Guénet, 2005; Yue et al, 2014). This suggests that even though the proteins involved in heat shock response are conserved, their regulation may follow distinct trends.

First of all, we will investigate whether orthologs have similar expression profiles throughout the response. Mahat et al. (2016) generated PRO-seq data during heat stress at various time points (2.5, 12, and 60 minutes) for mouse embryonic fibroblasts (MEF). For human data, although Vihervaara et al. (2017) had performed PRO-seq experiments on K562 (human myelogenous leukemia) cell line, it does not have the same kind of time course resolution as the MEF data. Therefore, we will generate the nascent transcription data for K562 cells at various time points during heat shock.

Hence, the heat shock protocol from Mahat et al. (2016) will be applied to K562 cells with and without HSF1 knockdown through RNAi in order to collect samples at various time points after heat shock. RNAi will follow the protocol from Östling et al. (2007) that also studied effects of HSF1 knockdown in K562 cells during heat shock. One adaptation is to also remove heat stress after 60 minutes by incubating the cells back at 37℃ and collecting samples while the cell recovers. Hence, the time points would consist of 0, 2.5, 12, 60 minutes after heat shock as well as 2.5, 12 and 60 minutes after its removal. The data after removing heat shock will be used in the analysis of recovery in Aim 3.

Next, following the PRO-seq protocol from Vihervaara et al. (2017), where they also applied it to heat shocked K562 cells, we will capture nascent transcription at the listed time points. In order to normalize the transcription level across the samples, ERCC RNA spike-in control will be added following the guidelines from Thermo Fisher™. Normalizing the read counts using spike-in control is important because there is likely a global increase or decrease of transcription during a heat shock response. Importantly, all the samples will have two replicates, and their correlations will be calculated.

With the generated data, genes will be classified into the aforementioned four classes (Fig. 2). For this part of the analysis, data from time points after removing heat shock are not used. Whether a certain gene is upregulated, downregulated, or unchanged is measured through DESeq2 analysis based on PRO-seq density as described in Mahat et al. (2016). Then, the first class can be distinguished by genes that are upregulated in normal K562 cells during heat shock but do not change expression under HSF1 knockdown. The second and third class will exhibit similar expression profiles with and without HSF1. However, those that show upregulation at 60 minutes but not at 2.5 and 12 minutes will be put in the third class. The remaining genes will fall under the fourth class.

Next, the comparative analysis between the derived classes of genes from K562 data and the classes from MEF can determine whether the genes follow a similar pattern of gene activation throughout a heat shock response. Since most of the protein-coding genes between human and mouse are orthologs with high conservation, the comparison is valid (Guénet, 2005). This comparison can reveal which proteins and their upregulation may be conserved between mammals, suggesting consistent roles in the heat shock response. To further explain their conservation, GO analysis will be utilized for functional enrichment. For other genes that have different expression profiles, further investigations to determine whether they have direct functions in the heat shock response may be necessary. In addition, for genes that are similarly regulated throughout the two species, the promoter region can be compared for enrichment of transcription factor (TF) binding motifs. Using CIS-BP database, a list of TF binding motifs derived from *in vitro* experiments will be curated. Then, the RTFBSDB tool will be utilized to

scan promoters for enriched potential heat shock related TF binding sites (Wang et al., 2016). With all the K562 data generated, analysis performed in Aim 1 can be applied again.

Aim 3. Transcription profile during recovery from a heat shock response.

The reason why the heat shock response is important to a cell is because it assists the cell to survive high temperature. Then, when the environment returns to normal condition, the cell must return to its normal gene expression as well. The activation of heat shock proteins after heat stress is quite rapid, so it raises the question whether the recovery phase occurs as quickly.

HSF1 activation and attenuation cycle is known to be regulated by levels of free HSP90 and HSP70 chaperone proteins. HSF1 is activated when chaperone proteins are occupied with properly folding damaged proteins. On the other hand, when proteostasis is reached, the freed heat shock proteins interact with HSF1 to convert them to an unactivated, monomer form. This negative feedback on HSF1 likely means that it has a role in the recovery for heat shock response. This motivates us to further test which genes are regulated in a HSF1 dependent manner.

To complement the recovery phase data generated from K562, the same kind of data for MEF will be created. Thus, the protocol from Mahat et al. (2016) will be adapted to capture recovering states. In the end, we will have a time course data during a heat shock response for both MEF and K562. Most of the analyses performed in Aim 2, can be applied to the current goal. We can similarly categorize the genes by HSF1 dependence of up- / down- regulation and their speed of recovery. Whether knocking down HSF1 with RNAi alters the expression profile implies the gene's dependence on HSF1. With the generated data, some hypotheses that can be tested are the following. First, is there a global change in transcription level throughout heat shock response? By normalizing the read counts with the spike-in control, we can test whether the global transcription rate increases or not. Second, is the downregulation of activated genes correlated with downregulation of TFs that activate it? Especially if genes that return to expression level during unstressed condition late have enriched binding sites of TFs that are downregulated quickly, we would be able to match TFs with their target genes. Thirdly, by monitoring the gene expression profile throughout recovery, one can time how quickly cells can actually return to normal.

Conclusion

Heat shock response is a dynamic, rapid and global change in the transcriptional program. Moreover, as it is a phenomenon existing in most clades of life, it harbors the basic principles of gene regulation. Through Aim 1, we attempt to decipher patterns in the genomic landscape that is correlated with differential regulation of genes during a heat shock response. Not only that, we study the hypothesis of whether divergently transcribed intergenic regions with HSF1 binding are enhancers activated by HSF1. Furthermore, the specificity of the heat shock response allows us to apply genetic approaches to parse dependency of regulators to gene activation, as in Aim 2. Lastly, the often overlooked facet of a heat shock response is its recovery. Through a time course PRO-seq data measuring nascent transcription after heat stress is removed, Aim 3 investigates the differential regulation of gene groups and the dependencies on HSF1. All in all, the proposal pursues a better understanding of the heat shock response and the basic gene regulation principle underlying it.

References

Baird, N. A., Douglas, P. M., Simic, M. S., Grant, A. R., Moresco, J. J., Wolff, S. C., ... & Dillin, A. (2014). HSF-1–mediated cytoskeletal integrity determines thermotolerance and life span. *Science*, *346*(6207), 360-363.

Battulin, N., Fishman, V. S., Mazur, A. M., Pomaznoy, M., Khabarova, A. A., Afonnikov, D. A., ... & Serov, O. L. (2015). Comparison of the three-dimensional organization of sperm and fibroblast genomes using the Hi-C approach. *Genome biology*, *16*(1), 77.

Dukler, N., Booth, G. T., Huang, Y. F., Tippens, N., Waters, C. T., Danko, C. G., ... & Siepel, A. (2017). Nascent RNA sequencing reveals a dynamic global transcriptional response at genes and enhancers to the natural medicinal compound celastrol. *Genome research*, *27*(11), 1816-1829.

Gidalevitz, T., Prahlad, V., & Morimoto, R. I. (2011). The stress of protein misfolding: from single cells to multicellular organisms. *Cold Spring Harbor perspectives in biology*, *3*(6), a009704.

Guénet, J. L. (2005). The mouse genome. *Genome Research*, *15*(12), 1729-1740.

Jamrich, M., Greenleaf, A. L., & Bautz, E. K. (1977). Localization of RNA polymerase in polytene chromosomes of Drosophila melanogaster. *Proceedings of the National Academy of Sciences*, *74*(5), 2079-2083.

Kelley, P. M., & Schlesinger, M. J. (1978). The effect of amino acid analogues and heat shock on gene expression in chicken embryo fibroblasts. *Cell*, *15*(4), 1277-1286.

Lemaux, P. G., Herendeen, S. L., Bloch, P. L., & Neidhardt, F. C. (1978). Transient rates of synthesis of individual polypeptides in E. coli following temperature shifts. *Cell*, *13*(3), 427-434.

Mahat, D. B., Salamanca, H. H., Duarte, F. M., Danko, C. G., & Lis, J. T. (2016). Mammalian heat shock response and mechanisms underlying its genome-wide transcriptional regulation. *Molecular cell*, *62*(1), 63-78.

Maza, I., Caspi, I., Zviran, A., Chomsky, E., Rais, Y., Viukov, S., ... & Hanna, S. (2015). Transient acquisition of pluripotency during somatic cell transdifferentiation with iPSC reprogramming factors. *Nature biotechnology*, *33*(7), 769.

McAlister, L., & Finkelstein, D. B. (1980). Heat shock proteins and thermal resistance in yeast. *Biochemical and biophysical research communications*, *93*(3), 819-824.

Östling, P., Björk, J. K., Roos-Mattjus, P., Mezger, V., & Sistonen, L. (2007). Heat shock factor 2 (HSF2) contributes to inducible expression of hsp genes through interplay with HSF1. *Journal of Biological Chemistry*, *282*(10), 7077-7086.

Richter, K., Haslbeck, M., & Buchner, J. (2010). The heat shock response: life on the verge of death. *Molecular cell*, *40*(2), 253.

Vihervaara, A., & Sistonen, L. (2014). HSF1 at a glance.

Vihervaara, A., Mahat, D. B., Guertin, M. J., Chu, T., Danko, C. G., Lis, J. T., & Sistonen, L. (2017). Transcriptional response to stress is pre-wired by promoter and enhancer architecture. *Nature communications*, *8*(1), 255.

Vihervaara, A., Duarte, F. M., & Lis, J. T. (2018). Molecular mechanisms driving transcriptional stress responses. *Nature Reviews Genetics*, 1.

Wang, Z., Martins, A. L., & Danko, C. G. (2016). RTFBSDB: an integrated framework for transcription factor binding site analysis. *Bioinformatics*, *32*(19), 3024-3026.

Yue, F., Cheng, Y., Breschi, A., Vierstra, J., Wu, W., Ryba, T., ... & Shen, Y. (2014). A comparative encyclopedia of DNA elements in the mouse genome. *Nature*, *515*(7527), 355.