Tentative Title: *Elucidating adaptive stress response pathways using transcriptomics*

Target Journal: ?????

Authors: X Y Z

**Key questions of this study:**

* Can we develop gene sets/signatures to identify adaptive stress response pathways?
  + Yes. While existing gene sets do exist, overlap and cross talk is likely more significant in these non-specific sets. As such consensus sequences can be compiled from these gene sets that are likely more typifying and exemplary of a stress state. How do these consensus sets score vs the existing?
* How do we evaluate the accuracy of these signatures?
  + GSEA signature scoring can be evaluated by ROC analysis using a hallmark chem set.
  + A maximized specificity and sensitivity unit of GS that work in concert can be constructed that optimizes scoring in an environment with a high degree of overlap
* Given the level of crosstalk between these pathways, are these gene sets sensitive and specific enough?
  + Yes – they operate as a unified stress classifier. Selectivity, sensitivity and discrimination were maximized at two stages. First when building consensus sets that eliminated significant amounts of cross talk and then again when using ROC/ AUC scoring to enhance discrimination.
* Limitations of this approach
  + Only as good as the definitional input – we defined stress at the beginning as a limited set
  + Only as strong as the Hallmark chemset – more should be used to create a training set and a classified set. Hard to do with a limited source. We should start building a database for this to expand our analysis
    - We did vary how we build the GS using external sources, building them from other data and not the profile data via a module methodology, so in essence we are comparing different sources
    - Should we include a the S1500 information as proof that if the size is reduced the efficacy is still ok? Similar to a resampling of a smaller gene space?

**Key outcomes**:

This is a baseline for improving the gene sets and/or scoring approach

* Consensus gene sets tend to score better than existing GS when compared suggesting our method for consensus gene set construction is sufficient
* A GS unit is better than using non optimized GSs
* Null target chemicals score the same as negatives and random nulls indicating that non stress agents would be ignored by this method
* DNA, HTS, ERS, HYP consensus sets have good accuracy and discrimination
* Surprisingly MTL profiles matched better with OXD GSs
  + This is possibly due to overlapping mechanisms or improper hallmarks – more needed
* OXD has some overlap with DNA which is expected

**List of key terms defined in this outline:**

**Canonical stress response pathways:** classical definition of stress response pathways based on chemical and environmental perturbance; these include DNA damage, metals stress, oxidative stress, unfolded protein/ER stress, heat shock stress, hypoxia, osmotic stress, the inflammatory response, and mitochondrial/metabolic stress.

**Source gene sets:** gene sets existing in the literature that are characteristic of or are built to characterize a stress response pathway

**Consensus Stress Response Gene Sets:**  gene sets built by selecting the most frequently associated member within existing GS describing a canonical stress system

**Hallmark chemset:** The set of characteristic chemicals that are typifying for a canonical stress state

**Stressome Signature Unit:** the final signature set composed of the top AUC scoring source and consensus gene sets designed to work as unit with which to classify stress response

Approach

1. **Defining the mechanics of stress:**

Stress is a reversable cellular condition resulting from external perturbances including mechanical and environmental damage (e.g., heat, UV), and chemicals disruptive to homeostatic cellular processes. Cellular stress response systems are the genes that make up biochemical pathways which allow cells to adapt to these perturbances and overcome adversity. Canonically, these stress systems are associated with DNA damage (DNA), protein misfolding stress (ERS), heat and cold shock (HTS), hypoxic stress (HYP), metal stress (MTL), and oxidative stress (OXD). Stress pathways have generally evolved to alleviate a singular condition but often share regulatory and response elements allowing for crosstalk and co-expression. As such it becomes important to isolate the central drivers and effectors of these systems to mechanistically characterize the exact systemic response that resulting in an adaptive outcome, wherein a cell reverses the effect of an insult or an adverse outcome wherein the cell is overcome by stress. Two components are then needed characterize a response, 1) a set of unique chemicals and agents inducing singular stress response systems, termed hallmark perturbagens, and 2) the genes that are most descriptive and central to a singular stress response, a gene set (GS).

\*Note: For the purposes of this study we will limit our definition to stress response, which is at the cellular level only, ignoring inflammatory and metabolic system responses that often include more complex interplay between cells and potentially secondary cell systems.

1. **Identify reference hallmark perturbagens and transcriptomic profiles to evaluate stress response gene sets**

Hallmark perturbagens were identified using three independent methods. An initial literature screen yielded 10 perturbagens associated with the canonical stress response pathways. This initial set was expanded to include chemicals identified in the Library of Integrated Network-based Cellular Signatures (LINCS) database. Small search sets (10 – 20 genes), were used query the LINCS database returning a ranked list of chemicals most associated with increased expression of these genes. Finally, the set was expanded once more with perturbagens found in the Comparative Toxicogenomics Database (CTD). Key genes linked with each canonical pathway were used to query the database resulting in a list of chemicals and perturbagens ranked by number of studies linking to the query gene(s*).*

* 1. Literature Screen
     + *An initial review of stress systems inducing perturbagens in the literature yielded a list of 20 well known chemicals linked to singular stress response system induction. The majority of the chemicals-stress system associations are described in Simmons, Ma, Chia, Lichtlen, Williams*.
     + *Core gene lists, that is the genes most centrally associated with each canonically stress system, were compiled during this review as well. Each core gene list ranged from about 10-20 genes that make up the central regulator, main transducers, response regulators, and effectors. Core genes were noted as up or down regulated using the best available information.*
  2. Searching for new chemicals based on signature matching (LINCS)
* *Small search sets (10-20 genes) were used to query LINCS. These search sets were filtered to be inclusive within the LINCS 1000 gene set and limited to up regulated genes only. Queries were conducted as “up regulated” via the LISTS and QUERY tools in in the CLUE dashboard.*
* *Returned LINCS DB responses were filtered to chemicals only and were ranked by median Tau score across all cell lines available.*
* *Chemicals were selected by inspection and LINCS associated targets.*

* *6 chemicals were selected when compared with literature and validated by other studies as having a characteristic association with a specific canonical stress response system.*

* 1. Chemicals identified by matching key genes (using CTD)
     + *Queries were performed in the CTD online search system and through (Imran/Grace’s system?)*
     + *Searches produced lists of perturbagens with known references to the query genes ranked by number of linking references as well as a language associated with the stress response system, the top 5 chemicals were selected*

All perturbagens compiled from the literature screen and CTD queries were searched against the NCBI’s GEO database yielding 25 transcriptomic data sets (profiles). These profiles, when combined with the 7 transcriptomic data sets taken from the LINCS GEO dataset form the hallmark chemset, a set of chemicals that typify the transcriptomic profile of each canonical stress response system. The hallmark chemset is composed of roughly 4 – 5 perturbagens per canonical stress category and has approximately 2-3 replicates per perturbagen. The DNA stress category has no replicates.

1. Construct adaptive stress response gene signature

Canonical stress response pathways have gene expression overlap and are not well represented among available gene sets on MSigDB or Harmonizome. Thus, a need for a purpose-built collection of gene sets characterizing canonical stress exists. Characteristics of this gene set collection must include non-overlapping genes and optimized discrimination between canonical stress response systems. This collection was built by first identifying existing gene sets available on MSigDB including those sourced from GO, Reactome, BioCarta, the MSigDB hallmark series, and others uploaded to MSigDB. Frequency of gene overlap between intra-group (same canonical gene set) and inter-group (all other canonical gene sets). An occurrence frequency metric was then calculated for each gene and consensus gene sets, sets composed of genes with the most intra-group frequency, of differential size were constructed at different size thresholds ranging from 50-477 genes. Each gene set (consensus or source) was scored against all other gene sets using receiver operating characteristic analysis. The tope gene set of each group was then taken and used to build a most discriminative stress response gene set.

* 1. Select TF and immediate targets based on literature
     + *Core gene lists, that is the genes most centrally associated with each canonically stress system, were compiled during this review as well. Each core gene list ranged from about 10-20 genes that make up the central regulator, main transducers, response regulators, and effectors. Core genes were noted as up or down regulated using the best available information*.
  2. To serve as a baseline, identify currently defined stress response gene sets from GO, Reactome or other sources ("known stress response gene sets")
     + *Currently defined stress response gene sets were identified first, by searching all MSigDB gene sets for a subset of central transcription factors, transducers, and effectors, selected from the core gene list for each canonical stress pathway. Queries contained the main TR/Response regulator and approximately 3 or more core effector genes. Queries were permutated until a maximum number of gene sets were identified*
     + *Identified gene sets were filtered by matching names and descriptions with the canonical gene set in question, e.g., DNA would require a “DNA” stress term in the title such as mismatch, or base excision repair.*
     + *Each canonical stress system sourced approximately 4-9 source gene sets, that is a set of gene sets that all have a functional association with a specific canonical stress response system*
  3. Create new consensus gene sets from MSigDB (DNA, ERS, HTS, HYP, MTL, OXD) based on merging existing stress system source gene sets and pruning to eliminate cross talk. This method produces a subset of gene sets that are “tuned” to work as a unit classifying stress response within a cross-linked transcriptomic environment.
     + *As stated above, all source sets were identified first by matching genes from each canonical stress response system core list with the contents of existing gene sets. These gene sets were compiled into a subset for each canonical stress system termed source gene sets*
     + *CONSENUS GENE SET MERGING: Consensus gene sets, gene sets that are composed of only the most source specific subset, were constructed using an occurrence frequency filter. Briefly: The occurrence of each gene within all canonical stress response source gene sets was counted and normalized for occurrence within a source set (termed intra-group occurrence) and within all other source sets (termed inter-group), this total abundance was then normalized to all occurrences within the gene universe. This is mathematically given as:*

*Where*

*S­fi = Occurrence Frequency of Gene i*

*GSia­ = Intra-group GS number, the total number of GSs within the inspected canonical GS pathway*

*GSit = Inter-group GS number, the total number of GSs in all of the source gene set groups less the number of GSs in the inspected canonical pathway*

*Ni, GSia = the number of occurrences for gene i within all of its intra-group GSs, GSia*

*Ni, GSit = the number of occurrences for a gene, i, within all of its inter-group GSs, GSit*

* + - *Once Sfi is calculated for each gene within each canonical gene set category (e.g., DNA, ERS, HTS, HYP, etc.), genes can then be rank ordered for their occurrence frequency, with genes having the highest occurrence frequency score within a canonical gene set being most central to that gene set. This assumes that major regulators will appear in most source gene sets of a specific canonical stress category, and, that effector genes, that are most diagnostic of a canonical stress response will be limited to lower frequency with in all other GS (inter group GSs) – as such a ranked consensus gene set can then be constructed for each canonical stress pathway.*
    - *CONSENSUS GENE SET PRUNING: once assigned a score, genes within each consensus gene set are rank-ordered by highest score. A threshold is then applied to constrain consensus gene sets to the top, n, scoring genes for each consensus subset. A threshold of 50, 100, 200, 300, 400, and 477(the maximum upper limit of all consensus sets) as applied. The result was a set of 36 consensus subsets with decreasing exclusion of scored intragroup genes.*
    - *A comparison of overlap of intragroup vs intergroup indicates that, following pruning, consensus sets had minimal overlap.*
  1. Evaluate overlap between gene sets to determine specificity. i.e. do the gene sets hit other pathways?
     + *Using a jacobian determinant (?) each 200 thresholded consensus GS was compared back to MSigDB identifying the most similar GS stored in the Database. Consensus GS returned expected GS from Database*

*Questions:*

*Do we need an index for cross/shared in part c*

1. Analyzing hits between signatures and transcriptomic profiles

Hallmark transcriptomic profiles were scored using myGSEA against the Consensus GSs and Source GSs. First all hallmark profiles were scored against all GSs and scored for accuracy using AUC following ROC analysis. The top AUC scoring GSs for each canonical stress pathway were compared against their scores in all other inter group canonical stress pathways. The winning intra-group/lowest inter-group scoring GS was compiled into a single stressome GS unit representing the best of GS for each . This unit has the benefit of having been built to operate as singlular system to achieve maximum scrutiny between itself and other GS within the unit leading to increased sensitivity. The stressome GS unit was then used in a second scoring analysis with the hallmark chem set, a null target dataset where in chemical targets with limited stress induction were selected, and a randomized null chem set, where chemicals with no relevant targets were randomized at the gene level. These results were characterized for accuracy by examining scoring depth to correct classification and null target/ randomized null score distributions.

* 1. Hallmark transcriptomics profiles were scored against source gene sets and consensus gene sets using myGSEA
     + *Hall mark chemicals were challenged with a signature list of all GS*
  2. Score specificity and selectivity of all GS using ROC/AUC
     + *Signature scores were analyzed for accuracy using pROC*
       1. *Binary scores assigned at each canonical stress category – 0 for control (neg) and 1 for case (positive)*
       2. *AUC values computed for all GS under all canonical stress pathways*
       3. *Ranked GS by AUC for each canonical stress pathway*
     + *Summarized AUC scores for all pathways indicating which pathways are best in each case and which work best as a unit*
  3. Build stressome unit
     + *top consensus stress gene set unit from top AUC consensus stress response categories / sum of AUC in all out-group sets*
  4. Analyze positive, negative, target null, and random null data GS scores
     + *A target null data set was constructed via the LINCs data set* 
       1. *Not stress activating targets were identified (e.g., estrogen, SSRIs, etc.)*
       2. *32 data points were taken and mixed with the HALLMARK chemcials*
       3. *The mixed dataset, hallmark+targetnulls, was rescored using the stressome unit*
       4. *ROC/AUC score were revaluated with the null targets included*
       5. *Success was evaluated by change in AUC*
     + *A random null was created by resampling gene expression values randomly within null target genes creating random expression profiles for null chemicals*
       1. *Random null samples were evaluated with the stressome unit\*\*\* NEED TO RERUN!?!?!*
     + *Analyzed score distributions for each positives (hallmark = test case), negatives, (hallmark ≠ test case), null target, and random null*
  5. Analyze accuracy of top consensus stress gene set unit with hallmark chemset
     + *Evaluated accuracy by scoring depth, that is, scoring depth necessary to reach a positive classification*
     + *Accuracy evaluated % of cases matching in the first or second score*