**Draft: Final Group Project:** **Should we take metformin to age gracefully?**

**First Draft to be discussed in Class Dec 3, Zoom Office hours on Dec 2nd or by request earlier.**

**Please read carefully before class and check on data availability.**

**Due date: Dec 17 by 5PM**

**Groups of 2-3. Each submitted text, graph, figure, table must be colored uniquely and unambiguously with each member of the team.**

**Introduction:**

Metformin is considered a “magical” drug because of its therapeutic potential to a myriad of diseases. It was originally developed to treat hypoglycemia in type 2 diabetes (T2D) patients but has also been shown to offer therapeutic benefit to other diseases including cardiovascular disease and cancer [5].

While the exact mechanism of metformin remains unclear, recently there has been an effort to determine whether it can be used for non-diseased states to promote “wellness.” A study found that a chronic low-dosage of metformin was able to extend the lifespan of mice, much like calorie-restriction extends lifespan.

However, no study has conclusively shown that metformin will have similar wellness-promoting effects in healthy humans and through what mechanisms. In this assignment, you will explore some of the relevant datasets to elucidate some of the wellness-promoting mechanisms of metformin and whether we would expect to see the wellness promoting effects in humans.

Biological clocks are a relatively new field that uses aging biomarkers such as methylation or others to predict biological vs chronological age. Feel free to explore it at your leisure.

**Data and Papers**

**MOUSE METFORMIN DATA FROM THIS PAPER:**

Metformin improves healthspan and lifespan in mice <http://www.ncbi.nlm.nih.gov/pubmed/23900241>

**HUMAN MUSCLE DATA FROM THIS PAPER:**

Increased SRF transcriptional activity in human and mouse skeletal muscle is a signature of insulin resistance [3] <http://www.ncbi.nlm.nih.gov/pubmed/21393865>

<http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE25462>

Diabetes, cancer, and metformin: connections of metabolism and cell proliferation. <http://www.ncbi.nlm.nih.gov/pubmed/22211893>

The target of metformin in type 2 diabetes <http://www.nejm.org/doi/full/10.1056/NEJMcibr1409796>

Cellular and molecular mechanisms of metformin: an overview [5] <http://www.ncbi.nlm.nih.gov/pubmed/22117616>

Biological Clocks Review

<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6520108/>

**HUMAN BRAIN DATA FROM THIS PAPER:**

https://www.ncbi.nlm.nih.gov/sites/GDSbrowser?acc=GDS5204

The goal of the project is to integrate everything you learned in class and try to **significantly improve** on your analysis in the previous homework (1-5).

* **Insulin Resistance and Diabetes in Muscle (Human)**
* **Brain Aging**
* **Metformin data in mouse**
* **ADDITIONAL DATA if you wish**

Format of submission: Paper that includes Abstract, Intro, Results, Discussion, Methods, Supplement (code + analysis results).

All writing or graphs must be colored by a unique color that is associated with one member of the team. Thus, analysis and writing must be done separately (but in collaboration).

Discussion with Professor in class and Zoom office hours is encouraged!

**Project Guidelines**

The following directions should serve as a guideline, but you shouldn’t hesitate to do additional analyses that you believe will provide useful insights. You will write a scientific report of publication quality in a typical Science or Nature style journal article (i.e. you should have a title, an abstract, introduction, methods, results, discussion, and references, and supplement). If you have any questions, feel free to ask.

**Data Retrieval and Normalization (probably not needed but for general reference)**

The first part of this project is to become familiar with retrieving data from a public repository (the Gene Expression Omnibus, or GEO) and properly normalizing it for additional analysis (if not done already).

Feel free to add any data you wish!

At this point, you can either 1) retrieve the raw data and proceed with normalization or 2) notice that these two studies already have the processed data available to you, thus you can just download the processed files and proceed to analysis. The processed data for all the samples in a study are usually provided in 3 different formats: SOFT, MINiML, and a Series Matrix File. Most of you will probably use the Series matrix file since it basically just a giant expression matrix with all the available phenotype data for every sample (e.g. BMI, age, gender, etc.). Make sure that the processed files are log2-transformed, which is a common data transformation for microarray analysis that you should do before additional analysis.

**Correlating Gene Expression to Phenotypes**

In the human study, since we have a reasonably sized cohort and several phenotypic measurements, we can ask which genes are correlated to certain physiological characteristics (phenotypes). Or whether they are differentially expressed in high range vs low range of the phenotype.

Because of the nature of the study (focus on aging / wellness), we recommend looking at the genes that are correlated with insulin resistance (log SI, aging or BMI). But feel free to choose other phenotypes.

**REQUIRED/RECOMMENDED ANALYSIS:**

1. Perform PCA, t-SNE or UMAP on the human data.
2. Plot a heatmap of the human data as informatively as you can.
3. Compute a correlation of genes in the human study to phenotypes. Identify the diagnostic genes.
4. Identify the differentially expressed genes for specific choices of phenotypes. Are they significantly different.
5. ideally plot selected box plots to show these differences.
6. Conduct some causal analysis!
7. Conduct Gene set enrichment analysis via DAVID or GSEA.
8. Dimension reduction, PCA or others (e.g. deep methods such as T-SNE) or UMAP can be used.
9. Clustering of the human data.
10. Classification of phenotypes using classification methodologies, INCLUDING **neural networks**, decision trees, Naïve Bayes, perceptron and comparison to other algorithms. Produce comparison and analyze errors.
11. Carefully analyze the Metformin data to identify genes responding to Metformin.
12. Are the biological processes affected by the drug associated with aging pathways, etc. (using David or GSEA).
13. Do you see meaningful similarities or differences between studies of aging.

Comments on GSEA/DAVID: Now that we have several lists of differentially expressed genes, we can ask which gene sets are overrepresented in these lists. Using your genes of interest, determine which gene sets they represent via DAVID and GSEA (DAVID ANALYSIS IS REQUIRED). Once again, if you are having trouble contact the instructors.

The gene sets you use should be the ones that you believe will be the most informative. We recommend looking at least at pathways (C2) and transcription factor targets (C3). For DAVID analysis check out KEGG pathways and GO (Gene Ontology).

**Data Exploration: Dimensionality Reduction and Principal Component Analysis AND MORE!**

Here, you will do some conventional or novel exploratory analyses on these data. Using the gene expression profiles for these studies, project the data onto PCA space using the 1st and 2nd principal component. PCA can be done in whatever programming language you are most comfortable with.

**Systems Biology of Human Disease Insights:**

Clustering, PCA, PCA projections, Set Intersections, Mapping to Networks (we can discuss), Network Diffusion. Be sure to very clearly state your hypotheses and your thought/design processes. For this part, you should use more than one dataset (required). We hope to discuss with teams their specific ideas and design.

You need to be very transparent about what you are doing for your analysis. Make sure you specify how you do analysis in the methods sections or supplements.