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

Due date: Dec 14, 2021 by 11:59 PM (extension only by submitting a partial report and requesting more time to do novel analysis).

Submission of Proposal with specific analysis planned Due by Dec 7th, 2021

Groups of 3 or 4.

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. So analysis and writing must be done separately (but in collaboration).

Discussion with Professor is encouraged!

**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 [4].

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.

**Required and Recommended Readings:**

**MOUSE DATA FROM THIS PAPER:**

Metformin improves healthspan and lifespan in mice [4] <http://www.ncbi.nlm.nih.gov/pubmed/23900241>

**HUMAN 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>

Diabetes, cancer, and metformin: connections of metabolism and cell proliferation. [2] <http://www.ncbi.nlm.nih.gov/pubmed/22211893>

The target of metformin in type 2 diabetes [1] <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>

**Project Guidelines**

We highly recommend that you start early!!!

Please read the papers first and take a look at the raw data and observe any challenges. We will post a clean version of the datasets as well. If you want to practice your own raw data analysis and compare to the posted version feel free to start raw analysis early (but expect challenges).

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). If you have any questions, feel free to ask the TA or instructor.

**Data Retrieval and Normalization**

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 !

You must to retrieve the datasets from the two previously mentioned studies by Martin-Montalvo et al. 2013 [4] and Jin et al. 2011 [3].

For convenience, you can find them directly below:

Metformin improves health span and lifespan in mice <http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE40936>

Increased SRF transcriptional activity in human and mouse skeletal muscle is a signature of insulin resistance <http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE25462>

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 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.). Note that the processed files are not log2-transformed, which is a common data transformation for microarray analysis that you should do before additional analysis. On a later date, we will provide you with cleaned, normalized data. However, we will take note if you start early and work with the unprocessed data to get a head start :)

**Correlating Gene Expression to Phenotypes, and Probe-to-Gene Conversion**

Now that we have downloaded these expression matrices, our next task is to ask what genes are differentially expressed between groups of interest (for example, 0.1% metformin-treated mice against control mice). The astute among you may have noticed a tool in GEO called GEO2R, which allows you to perform differential gene expression between simple comparisons. GEO2R basically runs an R script and the package ’limma’ to perform differential gene expression. However, it does not consider the given phenotypic parameters that you should account for (e.g. age, gender, etc.).

We suggest you read up on the Limma manual for how to take these into account. (THIS IS NOT REQUIRED). Feel free to use standard t-tests and conversion to Z-scores initially. Note that there are several different comparisons that we could make in these datasets, but for starters we recommend these comparisons:

1. 0.1% metformin-treated mice against control mice in muscle tissue
2. CR (Calorie Restricted) mice against control mice in muscle tissue
3. CR mice against control mice in liver tissue
4. 1% metformin-treated mice against control mice in liver tissue
5. FH− muscle against FH+ and T2D muscle

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.

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).

Note that you may need to condition on another variable (e.g. BMI) to get more sophisticated results, and we might cover this in a future class or lab discussion.

Finally, to convert the probe IDs to gene names, there are several ways to do this but the most convenient way we’ve found is to use Ensembl’s biomart or its associated R package, biomaRt.

**REQUIRED/RECOMMENDED ANALYSIS:**

1. Plot a heatmap of the human data.
2. Compute a correlation of genes in the human study to phenotypes
3. Identify the differentially expressed genes before and after treatment and ideally plot selected box plots to show these differences.
4. Conduct some causal analysis!
5. Conduct Gene set enrichment analysis via DAVID or GSEA.
6. Dimension reduction, PCA or others (e.g. deep methods such as T-SNE) or UMAP must be used.
7. Clustering of the human data.
8. Classification of metabolic phenotypes using classification methodologies, INCLUDING neural networks, perceptron and comparison to other algorithms.

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).

The actual GSEA approach is implemented as a standalone Java executable (and it’s also implemented in R). Both DAVID and GSEA will show you a variety of gene sets that you can use to do the enrichment analysis.

**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:**

Integrating multiple datasets to gain new biological insights is the best and most challenging part of the assignment, where we can finally put all our analyses together. One of goals of this course was to teach you how to do data integration, thus using all the knowledge and tools (clustering, classification, out of box thinking) you’ve gained in this course. We want to relate the longevity-wellness gene signature of metformin to the insulin-sensitivity signatures we see in human data. Any ideas to do this are welcome. Clustering, PCA, PCA projections, Set Intersections, Mapping to Networks (more info. will follow on this). Be sure to very clearly state your hypotheses and your thought/design processes. For this part, you should use all the data (required).

We hope to discuss with teams their specific ideas and design.

Technically speaking, you should correlate to log(SI). If you want to condition on other variables see partial correlations for more insights on how this is done.

You need to be very transparent about what you are doing for your analysis and state your hypothesis as cleanly as possible.

Make sure you specify how you do PCA OR ANY other analysis.

The key is to attempt to integrate the finding from the human study and the mouse study.

Also please state your hypotheses very clearly. We stated them only informally in class so please elaborate and make the study more precise.

**References**

[1] Ferrannini, E. The Target of Metformin in Type 2 Diabetes. New England Journal of Medicine 371, 16 (2014), 1547–1548.

[2] Gallagher, E. J., and Leroith, D. Diabetes, cancer, and metformin: Connections of metabolism and cell proliferation. Annals of the New York Academy of Sciences 1243, 1 (2011), 54–68.

[3] Jin, W., Goldfine, A. B., Boes, T., Henry, R. R., Ciaraldi, T. P., Kim, E.-Y., Emecan, M., Fitzpatrick, C., Sen, A., Shah, A., Mun, E., Vokes, M., Schroeder, J., Tatro, E., Jimenezchillaron, J., and Patti, M.-E. Increased SRF transcriptional activity in human and mouse skeletal muscle is a signature of insulin resistance. J Clin Invest 0, 0 (2011), 0.

[4] Martin-Montalvo, A., Mercken, E. M., Mitchell, S. J., Palacios, H. H., Mote, P. L., Scheibye-Knudsen, M., Gomes, A. P., Ward, T. M., Minor, R. K., Blouin, M.-J., Schwab, M., Pollak, M., Zhang, Y., Yu, Y., Becker, K. G., Bohr, V. a., Ingram, D. K., Sinclair, D. a., Wolf, N. S., Spindler, S. R., Bernier, M., and de Cabo, R. Metformin improves healthspan and lifespan in mice. Nature communications 4 (2013), 2192.

[5] Viollet, B., Guigas, B., Sanz Garcia, N., Leclerc, J., Foretz, M., and Andreelli, F. Cellular and molecular mechanisms of metformin: an overview. Clinical Science 122, 6 (2011), 253–270.