**Project description**

*What is topic of your research project / what is the research question? (max 5 lines)*

In this project, we profiled gene expression in three *in vitro* liver test systems (HepG2, iPSC‐HLC, and PHH) both with and without chemical exposure. In order to properly assess transcriptomic responses to these compounds, we first must characterize the baseline variability of different genes in the various test systems.

*What statistical question(s) do you want to pose to the students?*

What statistical methods and metrics best characterize the amount and sources of variability in baseline TempO‐Seq data? How can we adjust for this baseline variability in the following analysis?

*Please shortly describe the data source(s) available (number of subjects, number of*

*variables):*

PHH data: ~500 samples, ~3000 genes

HLC data: ~50 samples, ~3000 genes

HepG2 data: 125 samples, ~12000 genes

*If known already, would you like to perform particular statistical techniques on the*

*data (e.g, machine learning, mixed models, survival analysis, meta‐analysis)?*

Unsure as of yet.

*In what language is your dataset annotated (variable names and labels)?*

I can deliver the dataset with English annotation

**Content Questions**:

Can you please talk about how you collected the data? Also talk about the difference among genes?

(Can you briefly talk about your data? )

Why does HepG2 have more genes than the rest? (12k vs 3k) Are the 3k genes in other systems included in HepG2?

What does the numbers in the dataset mean? Do they have real numerical meaning or just labels? Labels means, for instance, using 0 and 1 to represent female vs. meal.  Is it numerical within some sort of range like 0-1? Is data from different systems rescaled or in the same range?

Were gene expression profiled in the three systems under the same conditions (e.g., time), or are there other factors we should consider adjusting for?

What is TempO‐Seq data ?

The three liver test systems have different numbers of samples and different numbers of genes. Can you briefly talk about the relationship between samples and genes? For instance, for the PHH data, does each sample have 3000 genes, and those 3000 genes are similar type of genes across different samples within PHH data? Then what about other data? Are **the genes in different data overlapped** to some extent?

Are there **any genes you want to pay more attention to or all the genes are treated to be the same**? (e.g., maybe there are some genes that are particularly interesting for your research topic but not for others).

======= Part =====

What kind of **follow-up questions** do you want to answer if you have adjusted for this baseline variability?

What **outcomes** do you **expect** from our analysis? There are two questions, which one would you give priority to?

Do you have data available already?  How was the data collected? Is there any known issues with the data, e.g., **missing values**?

**Clinents’ preference:**

Are there **specific software tools** you would like to use for analysis? Like, R, SPSS, Python or MATLAB.

How do you prefer to receive updates and results? (e.g., written reports, presentations)

How often would you like to meet or communicate throughout the project?

Do you have any additional questions or concerns?

We will send you over an e-mail about the agreement we reached today. Please let us know what you do not agree with.