BINF 531 Plan

**1. Data (better to have groups - will help visualizations)**

The datasets are...  
Species: Japanese quail,  
Life stage: Early life stage (in ovo),  
Tissue: Left-lobe of liver,  
Type: RNA-seq gene expression data (raw, annotated, and clean data all available, even including results after differential expression analysis)  
Design: Exposure to each of 8 chemicals, 1 control shared by 2 chemicals (4 batches), 2 doses for a chemical

|  |  |  |
| --- | --- | --- |
| **Chemical** | # ID | Date Tissues Collected |
| Ethinylestradiol (EE2) | 1 | March 2nd, 2017 |
| Benzo[a]pyrene (BaP) | 2 | March 30th, 2017 |
| Lead (Pb) | 3 | March 30th, 2017 |
| Selenomethionine (SeMe) | 4 | June 9th, 2017 |
| Chlorpyrifos (CPF) | 5 | March 2nd, 2017 |
| Fluoxetine hydrochloride (FLX) | 6 | June 9th, 2017 |
| Trenbolone (TB) | 7 | June 29th, 2017 |
| Hexabromocyclododecane (HBCD) | 8 | June 29th, 2017 |

2 chemicals were done at the same time and had one set for solvent control (DMSO).

Ethinylestradiol: birth control drug.  
EE has, albeit rarely (at the low dosages that are now used in COCs), been associated with cholestatic hepatotoxicity. Glucuronide metabolites of EE, via effects on the ABCB11 (BSEP) and MRP2 (ABCC2) proteins and consequent changes in bile flow and bile salt excretion, appear to be responsible for the cholestasis.[43] High concentrations of estradiol, via its metabolite estradiol glucuronide, are also implicated in cholestasis, for instance in cholestasis of pregnancy.[42] However, the incidence and severity of cholestatic hepatotoxicity appear to be much greater with EE than with estradiol, which is due to its 17α-ethynyl substitution and consequent reduced metabolism.[29]  
  
Benzo[a]pyrene: carcinogen. Its diol epoxide metabolites (more commonly known as BPDE) react and bind to DNA, resulting in mutations and eventually cancer. Souce: residential wood burning, charcoal barbecued food. Toxicity to nervous system, immune system, reproductive system.

Lead is a toxin that accumulates in soft tissues and bones, it acts as a neurotoxin damaging the nervous system and interfering with the function of biological enzymes, causing neurological disorders, such as brain damage and behavioral problems. Souce: Lead exposure is a global issue since lead mining and smelting, and battery manufacturing/disposal/recycling, are common in many countries. Lead enters the body via inhalation, ingestion, or skin absorption.  
  
Selenomethionine's antioxidant activity arises from its ability to deplete reactive oxygen species. Selenomethionine is readily available as a dietary supplement.  
  
Chlorpyrifos (CPS)  
organophosphate pesticide used to kill a number of pests including insects and worms. Exposure surpassing recommended levels has been linked to neurological effects, persistent developmental disorders and autoimmune disorders. Acute poisoning leading to a range of neuromuscular symptoms.   
  
Fluoxetine is used to treat depression, panic attacks, obsessive compulsive disorder, a certain eating disorder (bulimia), and a severe form of premenstrual syndrome (premenstrual dysphoric disorder).  
  
Trenbolone as trenbolone acetate, improves muscle mass, feed efficiency, and mineral absorption in cattle.  
  
Hexabromocyclododecane is a brominated flame retardant. may have adverse developmental and reproductive health effects.

*Japanese quail (Coturnix japonica) as a laboratory animal model:* “ For the past 50 years, the Japanese quail (Coturnix japonica) has been a popular animal model in numerous fields of research. The quail's 16-d developmental period and its easily accessible embryo make C. japonica a convenient model for studies of developmental biology. Because its lifespan is relatively short and its physiology is comparable to that of humans, the adult quail is useful for studies of aging and disease.”

**2. Work distributions**

project proposal (2-page?)

project presentation

analysis report (max 10-page)

**3. What steps to take**

Data understanding -

Dimension: original/raw dimension, cleaned dimension

Sample filtering: duplicated rows (summing up) & and then remove all the zero rows

data backgrounds

Data correlations - with pearson, spearman, etc

Data/variable filtering: Low variance, Low abundance

Data cleaning

Merging datasets - Correct batch effect (<https://academic.oup.com/biostatistics/article/17/1/29/1744261>)

Visualize data -

quality check,

outlier detections: box-plots, histogram

central tendency: mean, median, mode, IQR

distributions of the datasets

Data manipulation -

Normalizations: log transformation, auto-scaling, min-max scaling, etc)

Data analysis -

Heatmap

ANOVA (F-test → maybe we can do analysis based on groups of chemicals or dosage),

Testings: parametric and non-parametric testings - permutations, multiple testing corrections (bonferroni, FDR, etc)

1. **Unsupervised**

PCA (maybe try other dimension reduction methods and compare with PCA): scree-plots, loading plots, PC plots, quality of individual representation, contribution of individual to PCs, correlation circle,

Clustering - 1. Distance: single, complete, average group 2. Clustering methods: hierarchical, k-means, model-based (maybe decision tree later?)

1. **Supervised**

\*CV with MSE or RMSE

Regression - PCR, PLS regression, linear,

Classification - PLS, model performance (AUC, ROC), KNN

Data output