**Project Title:** Isoniazid metabolism inspired allele frequency and structural analysis

**Author**: Ava Xu

**Background**: Tuberculosis (TB) is a global public health epidemic that kills approximately 1.7 million people worldwide. Tuberculosis drug induced liver injury happens at 2-28% of time, is considered the most severe adverse effect that influences the current treatment with negative impact on patient compliance and treatment outcomes. One of the first line and most well-known TB agent, isoniazid, causes liver toxicity due to toxic metabolites produced in the liver, attributed to the variations in enzymes involved in this pathway like N-acetyltransferase 2 (NAT2), CYP2E1, and glutathione-S-transferases (GSTT1, GSTM1).

**Overview:** To study the effect of enzymatic activity in relation to allele frequency and structural changes from missense mutations. I constructed an integrated database that includes the allele frequencies of the variants for the enzymes participate in isoniazid metabolism pathways, structural annotation, clinical manifests and also mapping information with coordinates for visualization (Chimera). The goal is to analyze the results altogether looking at difference in allele frequency (how it differs in population genetics) and categorize if any patterns on the structures.

**Method**:

* Allele Frequencies: selecting ensembl gene transcripts for enzymes based on 3 filters, and output variant table, feed in variant table to variant effect predictor on ensembl
  + 1: with structure information
  + 2: missense mutations only
  + 3: sample has expression levels in the liver tissue
* Mapping: Use UCSF Chimera Visualization tool where you can define and color customized attributes by creating standardized txt files and specified by target residues
  + Goal is to compare common vs. rare variants and population differentiated variants on structure
* Annotation sources: combine ensemble, uniport, and uniport to protein database coordinates
  + Ensemble: ClinVar annotations
  + Uniprot: site annotation
* Analysis: Put these information together in guidance for hypothesis generation and mechanism insight
  + Population allele frequency
    - Overall AF, Maximal allele frequency difference
  + Annotations
  + Structure to Function
  + ddG trend (<https://structbio.vanderbilt.edu/~rinkerd/forAva/>)

**Results/Conclusion Summary:** details in final presentation

* Some population differentiated variants project patterns on the protein structures
* Only GSTM1 has statistically significant difference in ddG between common vs. rare variatns
* For NAT2, common variants reported with large ddG values, suggesting these variants potentially cause disruption of overall protein stability

Final presentation Link: <https://docs.google.com/presentation/d/1SdYwz8GbV5OoonDy6cgSk7UZuf59zsUeA9r3HAaI7pg/edit#slide=id.p>