**Statistical Design Consulting**

**SEMESTER REPORT**

**Spring 2025**

**Client:** Dr. Ying Li  **File Number:** 25-045

**Department:** Horticulture & Landscape Architecture **Major Prof:**

**Consultant:** Sumeeth Guda **Initial Meeting Date:** 03/26/25

**Meeting Attendees:** Dr. Ying Li, Sumeeth Guda, Dr. Chong Gu

**Statement of Problem:** Does incorporating chromatin data into a random forest model enhance the accuracy of the predictions of transcription factors for legumes.

**Goal of This Project:** Journal Article, Grant Application

**Background:**

The client is a faculty member in the Horticulture department who is investigating if gene regulatory networks are important for integrating CO2 availability and nitrogen supply in legumes. The background in the experiment is that increasing atmospheric CO2 concentrations has the potential to improve agricultural output. However, the photosynthetic gains from rising atmospheric CO2 concentrations are generally lower than the maximum predicted gains. This is caused by a down-regulation of photosynthesis that is connected to plant carbon-nitrogen imbalance when grown at elevated CO2. While this phenomenon occurs in legumes, it is not as severe due to their ability to exchange carbon for nitrogen with symbiotic soil bacteria. The client is interested in seeing how legumes sense, integrate, and respond to nitrogen and carbon/energy status through dynamic control of gene expression. And see how this enables the development of strategies for achieving maximum crop yield and quality in future climates. The client will build a gene regulatory network which combines gene expression data with epigenomic data measured across nitrogen and CO2 treatments. Their hypothesis for this experiment is that incorporating chromatin data will enhance the accuracy of the predictions of the transcription factors.

They came to the SCS to get consultation before for a grant proposal that is due on 3/28/25 as well as to gain the perspective of how a statistician would adjust the model parameters to account for the external Chromatin data.

**Progress During Current Semester:**

This project was resolved during the IM meeting. Dr. Gu suggested to the client to fix a single target gene and use different samples of that gene to create sample-specific weights. The openness of the TF gene could be used to modify the impact of the data, and if implemented as a multiplicative adjustment, this would mathematically be equivalent to modifying the predictor variable. In essence, chromatin data could be used to weigh the samples, transforming the approach into weighted least squares (WLS) regression. If GENIE3 allows for weight insertion, this method could be directly applied; otherwise, different versions of the weights could be created using histone data as an alternative strategy.

The client submitted the grant proposal, but it is possible they would want to continue with the SCS for the analysis phase of this project.

**Current Status: Continuing**