(Our project with Inari seeks to answer the question "what are the effects on the rest of the maize genes when we perturb a subset of the maize genes?" Essentially, we hope to construct a genetic network that can act as a look-up table to inform Inari of the genetic effects and side-effects of perturbing particular maize genes, as opposed to only observing unexpected effects later in the breeding process.)

In our literature review, we came across Connectivity Map (Subramanian et al, 2017), which achieved 81% accuracy when predicting the rest of the human genome from the 1,000 landmark genes, which serves as a good reference for our project which has similar goals but for maize. In comparison with CMap, our dataset has a few limitations, namely that 1. our dataset is not perturbation-drive, 2. our dataset size is significantly smaller, and that 3. the number of genes to be inferred is larger. As such, our EDA is done with the goals of ameliorating these limitations and investigating the feasibility of our task.

Before we started, we first preprossed the genetic expressions data by normalizing by gene length and by sequencing depth, log-transforming to reduce the disparity in scale. Addressing the third challenge, we removed 16.7% of the 46,430 maize genes that are not expressed in 80% of the samples and have the maximum expression value less than 2 TPM. We then performed principal component analysis (PCA) and t-Distributed Stochastic Neighbor Embedding (tSNE) and found distinct clusters of genetic expression profiles separated based on organism parts.

We shortlisted a series of potential methods of different algorithm classes (compiled by Saint-Antoine & Singh, 2019) including:

* Correlation-based: WGCNA; PGCNA
* Regression-based: TIGRESS; GENIE3; bLARS
* Information Theory-based: ARACNE; CLR; MRNET; PIDC

We also performed an initial naive approach to have multiple linear regression