# Incorporating molecular data-driven decisions towards soybean improvement

## 1. The problem I want to solve

Many agriculturally important traits such as yield, quality and some forms of disease resistance are controlled by many genes and are known as quantitative traits. It's labor intensive and time consuming to evaluate those traits. The strategy of genomic selection (GS) estimates the genetic merit of an individual based on molecular genetic information by simultaneously accounting for all DNA markers. GS is a form of marker-assisted selection that selects progeny lines or parents based on the estimated genomic estimated breeding values (GEBVs), which leads to shorter breeding cycle duration as it is no longer necessary to wait for late filial generations' performance trial. The problem I want solve herein is to build a genomic selection model for soybean complex traits (yield, protein and oil content).

## 2. Potential client and why do they care about this problem

The exponential growth of the world's population correlates to a stark rise in the demand for food production and an unprecedented opportunity for agribusiness companies. To meet this challenge, agribusiness companies are innovating for accelerating the release of superiority crop varieties. Their efforts take the form of artificial intelligence (AI) and remote sensors in the field, drones for crop monitoring and genomic selection for complex traits. So potential clients are agriculture companies like Monsanto, Sygenta and Bayer.

## 3. Data source information

GS usually needs a large amount of molecular markers covering the entire genome in term of all relevant genomic regions are represented by molecular markers and phenotypic data. The complete data set I will use is for 20,087 G. max and G. soja accessions genotyped with 42,509 SNPs is available for Wm82.a1 and Wm82.a2 in HapMap format. It's public available through the following link: https://soybase.org/snps/index.php. The phenotypic data is from a published paper https://www.g3journal.org/content/9/7/2253. Moreover, I will incorporate our own genotype and phenotype data (~1000 accessions)with the above big datasets.

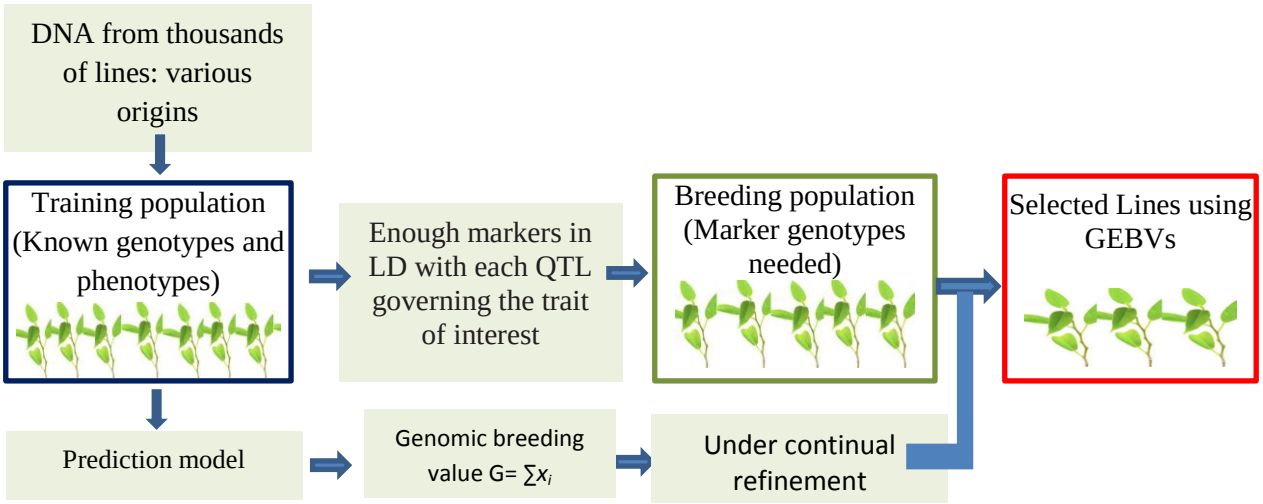
## 4. Brief outline I'll follow to solve this problem

Since the prediction model is calibrated with the genotypes and phenotypes of reference individuals constituting the training set, the following outline lists the major steps I will follow.

a) Collection of both phenotypic and genotypic data for each line in the training population (TP).
b) Integrating the genotypic and phenotypic data, conduct statistical analysis on TP to estimate allele effects at all loci on the phenotype.

c) Generation of a prediction model for the GEBVs that combines all the marker genotypes with their effects on the predictive value of each line.
d) Application of the prediction model on breeding population for which genotypes (but not phenotypes) are available. GEBVs are estimated and the best lines are selected for breeding.

The essential steps of GS are also shown as following:




**Flow chart showing the basic steps of genomic selection (GS).**

The standard linear model equation can be formulated as:

$$y = \sum_{j=1}^{n} x_{ij}\beta_j + e_i$$

where *y* is a phenotype value of the *i*th individual, $x_{ij}$ represents the genotype of the *i*th individual at the *j*th marker of 1 to *n* markers; $\beta_j$ is the allelic substitution effect at *j*th marker; *e* is the vector of random residual effects of ith individual. In *xi*, the allelic state of individuals can be coded as a matrix of 0, 1, or 2 to a diploid genotype value of AA, AB, or BB, respectively. Since the number of marker served as predictors (*p*) is usually far greater than the number of individual lines (*n*), it raise issues in multi-collinearity and overfitting among predictors. Basically, there are two regressions methods can be used, parametric and nonparametric regressions, including 11 models have been developed to address the "large *p* small *n*" issue. They are RR-BLUP (Meuwissen et al., 2001), least absolute shrinkage and selection operator (LASSO, Li et al., 2012) , elastic net (EN, Zou et al., 2005 ), Bayesian ridge regression (BRR, de los Campos et al., 2013), Bayesian version of LASSO (BL, Li et al., 2012) Bayes A (de los Campos et al., 2013), Bayes B (Heffner et al., 2011), Bayes C (de los Campos et al., 2013, 2016) and $\text{Bayes}_{C\pi}$ (Habier et al., 2011), reproducing kernels Hilbert spaces regression (RKHS, Gianola et al., 2008) and random forest (RF, Holliday et al., 2012).