Leveraging Genomic and Environmental Data for Prediction of Complex Traits in Wheat

de los Campos G.[1], Bogard M.[2] & Gouache D.[2]

[1]Michigan State University
[2]Arvalis Institut du Végétal
gustavoc@msu.edu

Most agronomically relevant traits are affected by large number of genetic (G) and environmental (E) factors. These factors can interact leading to genetic-by-environment interactions (G×E). Molecular markers (e.g., SNPs) and environmental covariates can be used to describe G and E factors. Modern genotyping platforms can produce genotypes at potentially hundreds of thousands of markers. Similarly, modern agro meteorological platforms allow measuring a large number of climatic variables; these can be integrated in ecophysiological models to derive hundreds of environmental covariates describing the conditions faced by the crop at different phenological stages.

In principle, markers and environmental covariates could be integrated into models that account for G, E and G×E. However, modeling interactions between predictors in two high-dimensional sets (e.g., molecular markers and environmental covariates) can be extremely challenging; indeed, if we only consider first order interactions, the number of possible contrasts needed to account for G×E equals the product of the number of markers and the number of environmental covariates.

In 2014, Jarquín et al. (TAG, 2014) proposed a methodology that allows integrating high dimensional genomic and environmental covariates in a model that accommodates G, E and G×E. The proposed approach is a multiple random effects model where: *(i)* the main effects of G and E factors are modeled using methods similar to those used in genomic selection (GS, Meuwissen, Hayes, and Goddard, Genetics 2001) and *(ii)* G×E is accommodated with a Gaussian process with a covariance structure which is a function of both markers and environmental covariates.

In the first half of our presentation we will describe the model proposed by Jarquín et al. (TAG, 2014) and present results (estimates of variance components and measurements of prediction accuracy) obtained by applying the model of to an extensive data set generated by Arvalis-Institut-du-Végétal, a French research and extension institute in agriculture. The data set (n=28,554) includes information from field trials (conducted from 1997 to 2014) where (601) wheat commercial varieties were tested in (243) different locations covering all the relevant agronomic regions of France. The varieties tested were genotyped with an SNP chip that rendered, after quality control, 213,339 SNPs. A total of 125 environmental covariates describing temperature, radiation and evaporative demand during five different phases of crop development were generated using climatic data and an ecophysiological model. We will present results on the proportion of variance explained by G, E and G×E and prediction accuracy in cross-validation.

One of the main objectives of Arvalis-Institut-du-Végétal is to produce recommendations as to what varieties are best fitted to each of the agro-climatic regions of France. We will demonstrate that sizable fraction of the environmental variance (about 50% of the phenotypic variance of grain yield) can be attributed to year-location

interactions (i.e., within-location year-to-year differences); this, together with the fact that G×E makes a substantial contribution to grain yield, suggests that rather than predicting the average performance of varieties in target locations, we may need to predict the distribution of yield of any candidate variety over possible realizations of the environmental conditions within a target location. Unfortunately, even with the extensive network of trials established by Arvalis-Institut-du-Végétal, not all varieties are tested in all target locations. Moreover, and perhaps more importantly, for any given location, data cover only a handful of years. These data are not sufficient to derive predictions of the expected distribution of yield over possible realizations of the environmental conditions. In the second half of our presentation we will discuss how we can integrate historical weather data and data from field trials to predict the distribution of yield of candidate varieties in target locations.