Jinliang Yang | March 2016

Research Statement

My resesarch focuses on bridging the gap between genotypes and agronomically important phenotypes. My current research takes quantitative and population genetic approaches to dissect the genetic architecture of several yield-related traits to explore the genetic basis of heterosis and predict hybrid performance using information based on patterns of evolutionary constraint. I have developed experimental and computational methods to map QTL (**XP-GWAS**), to incorporate biological information into Genomic Selection, to impute heterozygote genotypes for a pedigree-based GBS data (**imputeR**), and to manipulate large SNP data (**zmSNPtools**). The above tools or packages are publicly available on my Github account (*https://github.com/yangjl*). At a broader scale, I am keen to integrate various large-scale biological data sets such as phenomics, genomics, transcriptomics, methylomics data and functional annotations to boost the power of Genomic Selection. In the sections below, I describe these areas of research and their future directions.

Genetic Basis of Heterosis

Background

Hybrid vigor or heterosis is of substantial importance to crop improvement but its genetic basis remains controversial. Population genetic theory has long argued that heterosis can be explained largely by complementation of recessive deleterious alleles, but this model is inconsistent with evidence from polyploid hybrids that suggests an important role for gene dosage. A general model is required to resolve this controversy and enhance the heterosis prediction for future breeding.

Incomplete dominance explains heterosis

During my PhD work, using a GWAS approach, I found the level of heterosis of different traits correlated with the number and the magnitude of effects of positive dominant gene action (**Figure 1**, Yang et al., in preparation). During my postdoctoral work, I have shown that a model of incomplete dominance both fits predictions of population genetic theory and explains empirical results from polyploids. I tested this model in yield trials of a partial diallel population of elite maize inbreds. We re-sequenced the genomes of all 12 parents of the population, annotated variants as putatively deleterious using a measure of evolutionary conservation. I then estimated the degree of dominance of each variant, and by using an identity-by-descent approach to perform genomic prediction on

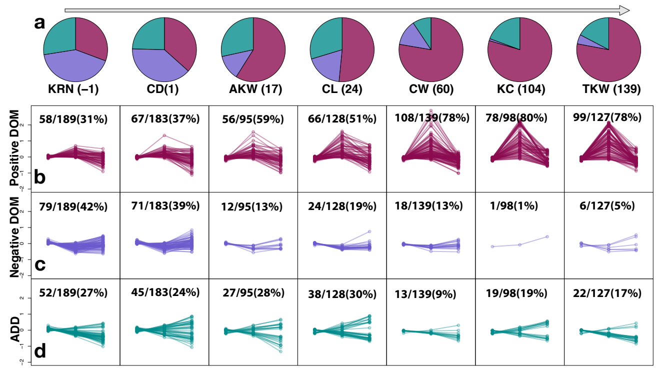

*Figure 1: **Gene action for individual trait-associated variants (TAVs).** In panel (a), the pie charts show the proportion of TAVs exhibiting positive dominance, negative dominance and additive gene action. In the below panels, TAVs exhibiting positive dominance (b), negative dominance (c) and additive (d) gene action were plotted separately. On the axis of each plot, magnitude of effects for B73 alleles, heterozygotes and non-B73 alleles. The seven traits were ordered by their levels of heterosis (HPH).*

haplotype blocks, show that our model predicts heterosis for grain yield and other traits better than simple models of pure additive or pure dominance (**Figure 2**, Yang et al., in preparation). I also show that *in silico* predictions of triploid phenotypes from our data are consistent with empirical observations in polyploids. These results thus unite two disparate hypotheses for the genetic basis of heterosis and paves the way for the use of Genomic Selection models to better estimate and predict this agronomically important phenomenon.

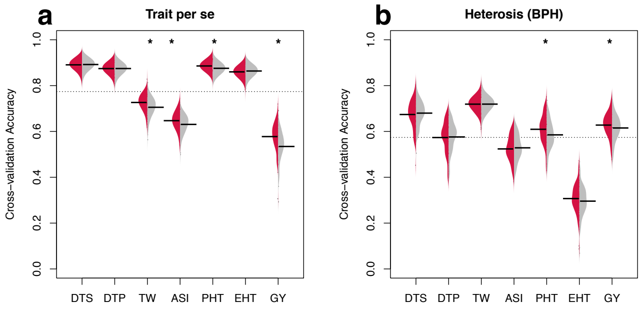

*Figure 2: **Genomic prediction models.** Beanplots represent prediction accuracy estimated from cross-validation experiments for traits per se (**a**) and heterosis (**b**). Prediction accuracy using estimated dominance values for each SNP is shown on the left (red) and permutation results on the right (green). Horizontal bars indicate mean accuracy for each trait and the gray dashed lines indicate the overall mean accuracy. Stars above the beans indicate significantly (permutation FDR < 0.05) higher than cross-validation accuracy.*

Future Directions

In the future, I will continue work on heterosis by using larger, broader, or even polypoid populations. These data allow opportunities to dissect the controlling loci to finer regions using GWAS approaches. I will conduct experiments to precisely estimate the contribution of different modes of inheritance and to study how evolution/domestication/selection shapes the change of mode of inheritance. In addition, I will study the functional aspects of heterosis by exploring its relationship with better annotation information, recombination, gene expression and DNA methylation at the population level.

Genetic Architecture of Agronomically Important Traits

Background

Genomic Selection emerged as a competing technology for crop improvement, which allows breeders to predict plant performance without extensive field evaluation and advanced to next cycle of selection more rapidly. The success of Genomic Selection relies on better understanding the genetic architecture of complex traits, in particular modes of inheritance, the number of loci controlling traits, and the distribution of their effects.

Dissecting QTLs for Yield Related Traits

I have conducted GWAS and dissected the genetic architectures for seven yield-related traits (Yang et al., in review, Sosso et al., 2015), six shoot apical meristem (SAM) related traits (Leiboff et al., 2015) and two root-related traits. I have also conducted genetic validation experiments with three different GWAS approaches using a highly heritable trait, kernel row number (KRN). Genetic validation results showed that approximately 50% of trait-associated variants exhibited associations with the KRN trait in at least one unrelated population. Importantly, about 60% of the trait-associated variants exhibited associations in only one of the three statistical approaches. This finding demonstrates that the three GWAS approaches are complementary. I also found trait-associated variants in low recombination regions were more likely to exhibit associations in independent populations than in regions of active recombination, probably as a consequence of linkage disequilibrium (Yang et al., in review).

In addition, we sought to characterize the genetic basis of gravitropism of maize seedling roots using machine vision (**Figure 3**). In this study, I showed that 8% of the variation in root angle in the resulting data set was explained by the phase of the moon at the time of measurement. I confirmed this phenomenon

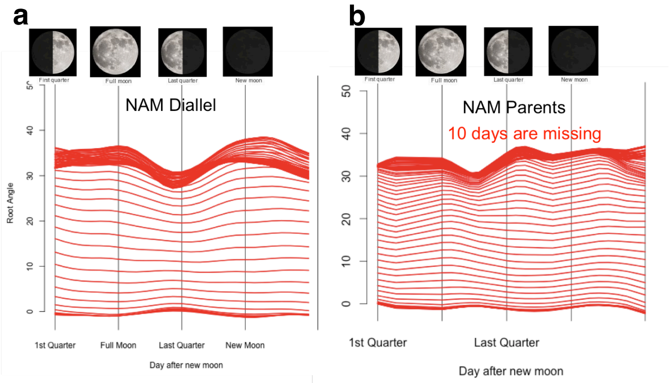

Figure 3: ***Relationship between root angle and lunar phase.*** *BLUE values for the lunar phase effects were plotted for the 61 time points (red lines) observed using a set of NAM diallel (**a**) and NAM parental lines (**b**). Lines are smoothed using 'loess' function in R.*

in an experiment designed to test for associations between gravitropism and the lunisolar tidal force and identified genotype-specific responses. These results suggest that the lunisolar tidal force may affects the gravitropic response of maize seedling roots. We are doing more experiments to validate this finding (Yang et al., in preparation).

Future Directions

With extensive QTL and association mapping experiences, I will continue work on the genetic dissection of complex traits. I will focus on fine mapping and molecular characterization of chromosomal regions controlling for yield-related traits. I will develop suitable genotyping techonologies for my experiments, i.e. GBS, sequence-capture or RNA-seq based SNP calling. I will collaborate with other researchers to utilize emerging technologies such as machine vision and image analysis for high-throughput phenotyping. In addition, I will integrate multiple biological big data to assistant QTL and association mapping.

Genetics and Genomics Methods Development

Background

In IT world, "90% of the data in the world today has been created in the last two years alone". This causes the "big data" challenges. The situation is very similar for biological big data. For example, the maize community alone deposited $>$ 20 Tera base-pairs into the NCBI short reads archive (SRA) database in the past two years. Until the end of 2014, $> 2,300$ Tera bases reside there for all species and most of them are not fully analyzed. The biology community needs tools to process and analyze large biological datasets if we hope to exploit these data to advance the biological understanding and crop improvement. I am keen to develop and share genetic and genomic tool sets to serve the community.

XP-GWAS, imputeR and SNP tools

During my PhD work, I had developed a method (called **XP-GWAS**, extreme phenotype GWAS, `https://github.com/schnablelab/XP-GWAS`) for conducting GWAS that does not require genotyping of large numbers of individuals (Yang et al., 2015). This method measures allele frequencies in pools made up of the phenotypic extremes of a population, enabling the discovery of associations between genetic variants and traits of interest. XP-GWAS was able to resolve several linked QTL and detect trait-associated variants within a single gene under a QTL peak. XP-GWAS is of particular value for plant species that do not have access to extensive genotyping resources, such as the wild progenitors of crops, orphan crops and other poorly characterized species.

Recently, I developed an R package (called **imputeR**, `https://github.com/yangjl/imputeR`) to impute missing data and to correct the inaccurate heterozygote SNP calls from raw genotyping-by-sequencing (GBS) data in a pedigree design. I have applied this package to impute two data sets, a teosinte population and a maize landrace population of ∼5,000 progeny each. The imputation results have low error rates and have high recovery of missing data, and that it outperformed industry standard methods of imputation by incorporating Mendelian segregation information. I hope the package may help to broaden the usage of GBS by improving its SNP calling quality, especially for reducing the error rates for heterozygote SNP calls. I also shared a collection of perl, python and R codes that I have accumulated in my research for various ways of SNP manipulation (**zmSNPtools**, `https://github.com/yangjl/zmSNPtools`).

Future Directions

I will continue this path of computational tools development in the future. My area of interests will be focusing on processing genotyping and sequencing data, RNA-seq and whole-genome bisulfite sequencing analysis, and especially on statistical methods development for association mapping and Genomic Selection.