Systemic Analysis of Biological Data from an Isogenic Maize Line

Merritt Bryer Burch, Boris Shmagin, Vivek Shrestha, Donald Auger

Department of Biology and Microbiology, South Dakota State University, Brookings SD 57006, USA

Introduction & Background

Doubled-Haploids (DH)

- Monoploid plants can chemically or naturally diploidize to create completely homozygous progeny so that both parent and progeny are identical genotypically and phenotypically
- DH maize is useful in creating inbred lines for breeding programs to produce hybrids
- Sprague et al., (1960) demonstrated that polymorphisms in quantitative traits among DH maize emerged at a faster rate than the known rate of discrete mutations
- Genetic breakdown of inbred-stocks could be detrimental to breeders, farmers, and researchers

Principal Component Analysis (PCA)

- PCA is a statistical technique used to summarize systematic patterns of variation into a small number of subsets, called principal components
- These principal components are combinations of measured traits whose variability explains a majority of the relationships in our data
- These analyses are useful when attempting to explain patterns in a population based on multiple variables, unlike single variable statistical techniques like 1/2-way ANOVAS or mixed effect models

Materials and Methods

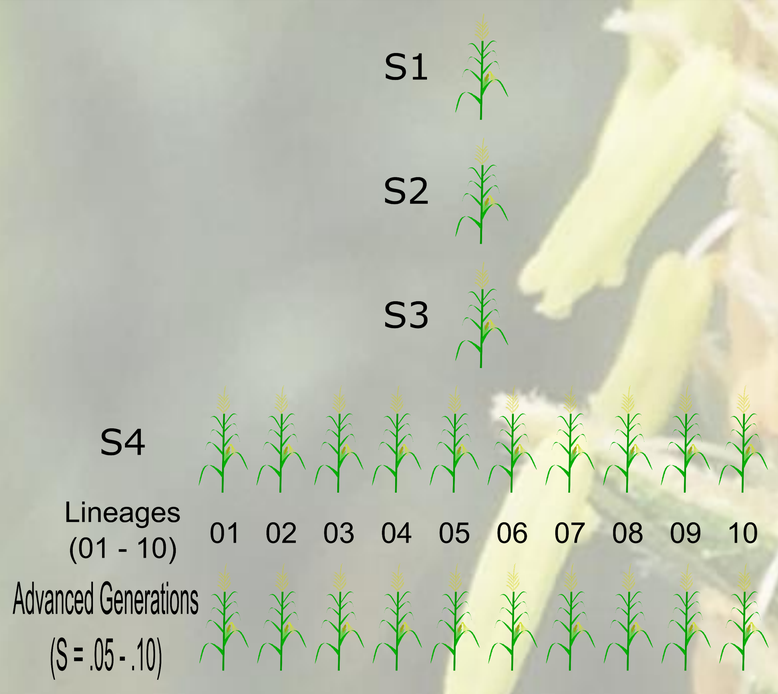

Fig. 2: A single-seed descent line created from a B73 doubled-haploid plant was sequentially self-pollinated. Generation 3 (S3) seeds were randomly selected to create ten separate DH lineages. These lineages were advanced and assessed for quantitative traits.

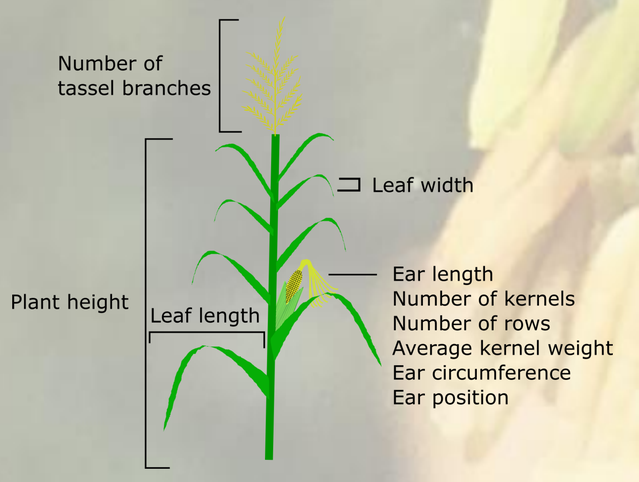

Fig 1: A typical maize plant showing the location of ten quantitative traits measured. A total of thirteen traits were measured and tested

- In the summer of 2014 all ten lineages, with two or more generations, were grown at SDSU in randomized blocks.
- The data were then centered and scaled before PCA
- Factor analysis on was done on five dimensions with an orthogonal rotation
- Data visualization and analysis was performed in the psych, ggplot2, and factoextra packages in R Studio (V: 3.3.2)

Results

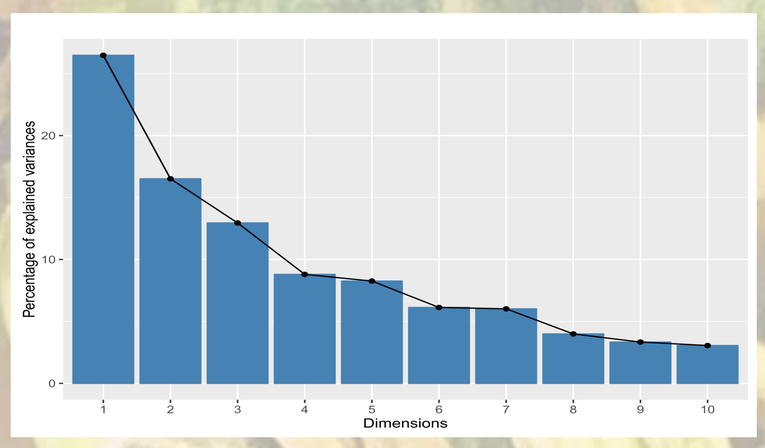

Fig 3: Scree plot showing the percentage of explained variances and the number of dimensions. A total of 5 dimensions, equivalent to 73% of the variance would be sufficient in explaining the data.

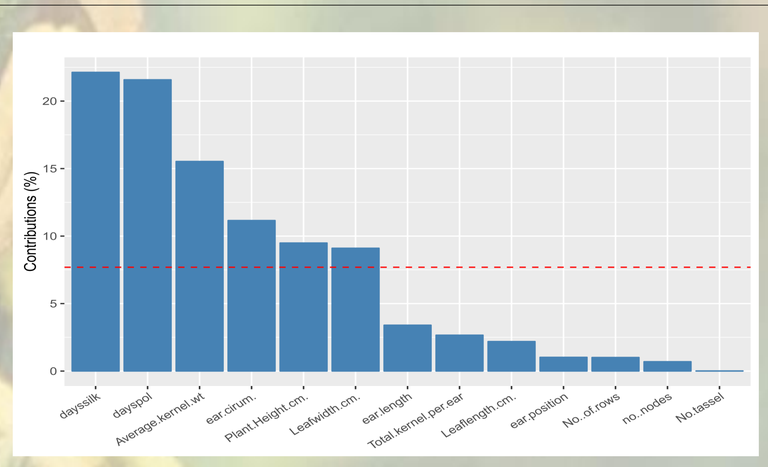

Fig 4: Bar chart showing the percent variance contribution of each trait to Dimension (principal component) 1. Dimension 1 accounts for 27% of the variability in the data. The days to silk and pollination, along with the average kernel weight are extremely important in explaining the data.

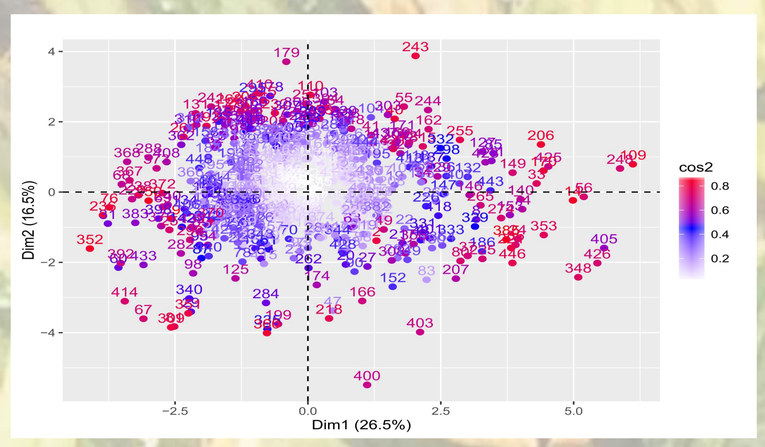

Fig 5: A factor map showing the relative importance of each individual in the first two dimensions of the PCA. In this map, no true clustering is observed, potentially due to this data being collected on genetically similar plants. Influential observations correspond to some sequentially self-pollenated DH lineages.

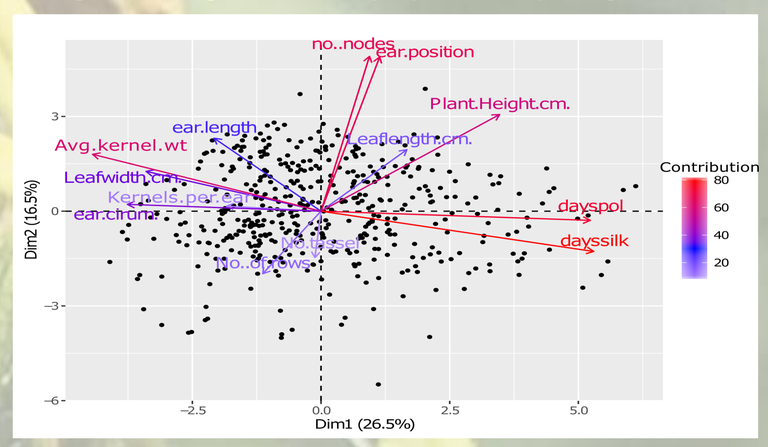

Fig 6: A biplot displaying our traits and their relative contribution in explaining the data's variance in the first two dimensions along with the individuals. Positively correlated variables are grouped together, negatively correlated variables are on opposite sides of the plot origin, and the distance between variables measures their quality.

Loadings	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5
Leaf width	-0.59				
Avg. kernel wt.	-0.87				
Days to pollen	0.89				
Days to silk	0.90				
Leaf length		0.59		-0.42	0.50
Plant height	0.36	0.71			
Ear position		0.72		0.41	
No. nodes		0.81			
No. of rows			0.79		
Total kernels			0.76	0.46	
Ear circum.	-0.43		0.66		
Ear length				0.82	
No. tassels					0.85
Proportion var.	0.24	0.17	0.14	0.11	0.08
Cumulative var.	0.24	0.40	0.54	0.65	0.73

Table 1: Orthogonal factor analysis loadings showing the association of our variables to each factor. Loadings are interpreted as normal regression coefficients. Unrotated factors had weaker correlations.

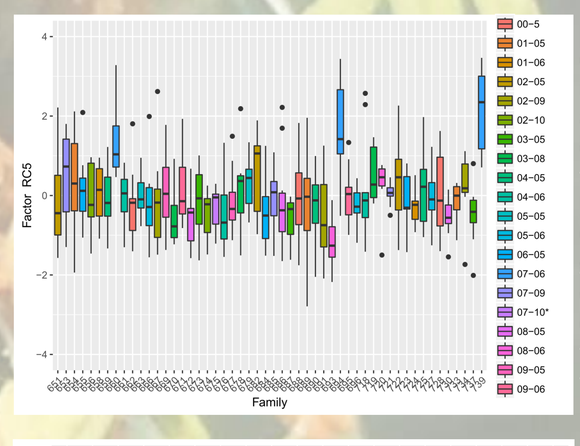

Fig 8: Box plot depicting the median and the 25th and 75th percentiles of Factor 5, the field family, and the lineage-generation. One lineage-generation, 07-06 consistently stands out for a certain trait, the number of tassel branches.

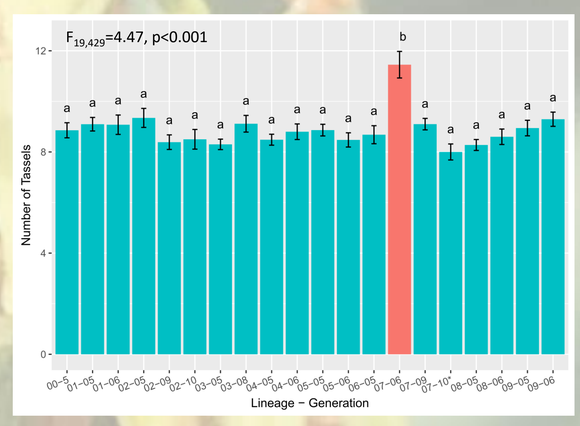

Fig 9: Bar chart depicting the mean and standard error of the number of tassel branches and the lineage-generation. Least squared means post-hoc results show significant relationships that have unique letters.

Conclusions

- Biplots showed that days to silk and pollen emergence, and average kernel weight were the most important in our data
- Reduction of our data to these few traits for future analyses would be useful in explaining patterns in our population as a whole
- Explaining individual lineages, however, and their heritable changes in advanced self-pollinated generations might not work in PCA or other multivariate analyses
- No clustering of individuals within traits is observed in Fig 5 & 6, suggests that our DH data may have too little variance to find any large relationships
- Factor analysis helped to correlate the data more efficiently and found previously unseen trends in the data (Fig. 8 & Fig 9)

Future Work

- Perform RNAseq on lineages that show significant, heritable changes in quantitative traits to determine genetic and epigenetic influences on DH genes
- Do PCA on partial study replicates from 2015 and 2016
- Do cluster analysis of our traits and lineages

Acknowledgements

Doubled-haploid plants were produced by Akio Kato and provided by Jim Birchler (University of Missouri)

This research was funded by the SDSU Agriculture Experiment Station