Data driven trait quantification across a maize diversity panel using hyperspectral leaf reflectance

Michael Tross[1], Marcin Grzybowski[1], Aime V Nishimwe[1], Guangchao Sun[1], Yufeng Ge[1], and James C Schnable[1]

[1]University of Nebraska-Lincoln

November 22, 2022

Abstract

Scoring plant phenotypes across large populations in multiple environments is a necessary precondition to both using natural genetic diversity to build genotype to phenotype models, study genotype by environment interactions and to carry out plant breeding to develop high yielding and more resilient cultivars. Here we explore data driven approaches using latent representations of leaf reflectance data collected from a large field experiment consisting of a subset of diverse maize lines drawn from the Wisconsin diversity panel (Mazaheri et al., 2019). In this experiment, 2 replicates of 752 inbred lines from the Wisconsin diversity panel were grown in field conditions. An ASD spectrometer was used to collect data on intensity of light reflected by leaves at 1 nanometer wide intervals between350 to 2,500 nm, resulting in a total of 2,151 reflectance intensity values measured for each plot. Two dimensional reduction approaches were evaluated for this dataset: conventional principal component analysis and an auto-encoder based neural network. Ten principal components were sufficient to summarize 99% of variance in the dataset. An autoencoder neural network comprising of an encoder having three dense layers and a decoder having four dense layers was able summarize variation within the dataset at a validation loss of 0.006 using 10 latent variables. A number of principal components and latent variables were correlated with several phenotypes quantified for a subset of the same field grown research plots (Figure 2A;2C). Chlorophyll, the major photosynthetic pigment in plant leaves, plays a substantial role in determining the overall pattern of reflectance for maize leaves. The abundance of chlorophyll was significantly correlated with PC2 (R2 = 0.31) (Figure 2B) which explained 11% of the total variance in higher spectral reflectance data. However, autoencoder based summary of the same trait dataset appears to have more accurately captured variation in chlorophyll abundance within this field trial with LV8 exhibiting a R2 = 0.59 (Figure 2D) with ground truth chlorophyll measurements. Both PCA and autoencoder based dimensional reduction captures a mix of variables which were heritable (i.e. a large proportion of total variance was attributable to differences between genotypes) and variables that were not heritable. Two of ten PCs evaluated exhibited H2 values >0.5 as did four of ten latent variables generated (Figure 3A; 3B). Genome wide association studies (GWAS) conducted using high heritability principal components and latent variables identified significant signals in 2 out of 6 cases (Figure 4A; 4B). Ongoing work is needed to evaluate the potential of using candidate genes underlying GWAS peaks to assign putative biological roles to latent variables estimated from raw sensor data by autoencoders or other dimensional reduction approaches.

Data driven trait quantification across a maize diversity panel using hyperspectral leaf reflectance

Michael C. Tross[1], Marcin Grzybowski[1], Aime V. Nishimwe[1], Guangchao Sun[1], Yufeng Ge[2] and James C. Schnable[1*]

[1] Quantitative Life Sciences Initiative, Center for Plant Science Innovation & Department of Agronomy and Horticulture, University of Nebraska-Lincoln, Lincoln, NE, 68588 USA
[2] Center for Plant Science Innovation & Department of Biological Systems Engineering, University of Nebraska-Lincoln, Lincoln, NE, 68588 USA
[*]schnable@unl.edu

Scoring plant phenotypes across large populations in multiple environments is a necessary precondition to both using natural genetic diversity to build genotype to phenotype models, study genotype by environment interactions and to carry out plant breeding to develop high yielding and more resilient cultivars. The collection of plant phenotype data from the field has traditionally been a manual and labor intensive process with limiting economies of scale. As labor costs have increased and the need for increasingly large and comprehensive datasets has grown substantial investments have been made in automated data collection, e.g. by ground or air based automated systems, fixed imaging stations, and satellites as well as human carried sensor platforms that can rapidly capture large amounts of data from individual plants or plots. The raw sensor data generated by any of these approaches must be processed via various algorithms or trained models to extract numerical estimates of traits of interest.

Hyperspectral reflectance data is a model for the challenges of high throughput plant phenotyping. Hyperspectral reflectance measurements can be captured with a range of sensor technologies and multiple studies have demonstrated how the raw data generated by these sensors can be employed to train models for estimate a range of plant traits including potassium, chlorophyll, nitrogen, specific leaf area and phosphorus ((as reviewed (Grzybowski et al., 2021)). However, as with many model training approaches to processing sensor data generated from plant phenotyping efforts, the capacity to train models to quantify new traits is typically bottlenecked by the generation or availability of large and high quality ground truth datasets for the traits of interest with which to train and evaluate the models.

Genetic variation among plant varieties has been quantified using latent phenotyping. Previous approaches using this method have used RGB and LIDAR data without ground truth measurements (Gage et al., 2019; Ubbens et al., 2020). Here we explore data driven approaches using latent representations of leaf reflectance data collected from a large field experiment consisting of a subset of diverse maize lines drawn from the Wisconsin diversity panel (Mazaheri et al., 2019).

In this experiment, 2 replicates of 752 inbred lines from the Wisconsin diversity panel were grown in field conditions. An ASD spectrometer was used to collect data on intensity of light reflected by leaves at 1 nanometer wide intervals between350 to 2,500 nm, resulting in a total of 2,151 reflectance intensity values measured for each plot. Two dimensional reduction approaches were evaluated for this dataset: conventional principal component analysis and an auto-encoder based neural network. Ten principal components were sufficient to summarize 99% of variance in the dataset. An autoencoder neural network comprising of an encoder having three dense layers and a decoder having four dense layers was able summarize variation within the dataset at a validation loss of 0.006 using 10 latent variables.

A number of principal components and latent variables were correlated with several phenotypes quantified for a subset of the same field grown research plots (Figure 2A;2C). Chlorophyll, the major photosynthetic pigment in plant leaves, plays a substantial role in determining the overall pattern of reflectance for maize leaves. The abundance of chlorophyll was significantly correlated with PC2 ($R^2 = 0.31$) (Figure 2B) which explained 11% of the total variance in higher spectral reflectance data. However, autoencoder based summary of the same trait dataset appears to have more accurately captured variation in chlorophyll abundance within this field trial with LV8 exhibiting a $R^2 = 0.59$ (Figure 2D) with ground truth chlorophyll measurements.

Both PCA and autoencoder based dimensional reduction captures a mix of variables which were heritable (i.e. a large proportion of total variance was attributable to differences between genotypes) and variables that were not heritable. Two of ten PCs evaluated exhibited H2 values >0.5 as did four of ten latent variables generated (Figure 3A; 3B). Genome wide association studies (GWAS) conducted using high heritability principal components and latent variables identified significant signals in 2 out of 6 cases (Figure 4A; 4B). Ongoing work is needed to evaluate the potential of using candidate genes underlying GWAS peaks to assign putative biological roles to latent variables estimated from raw sensor data by autoencoders or other dimensional reduction approaches.

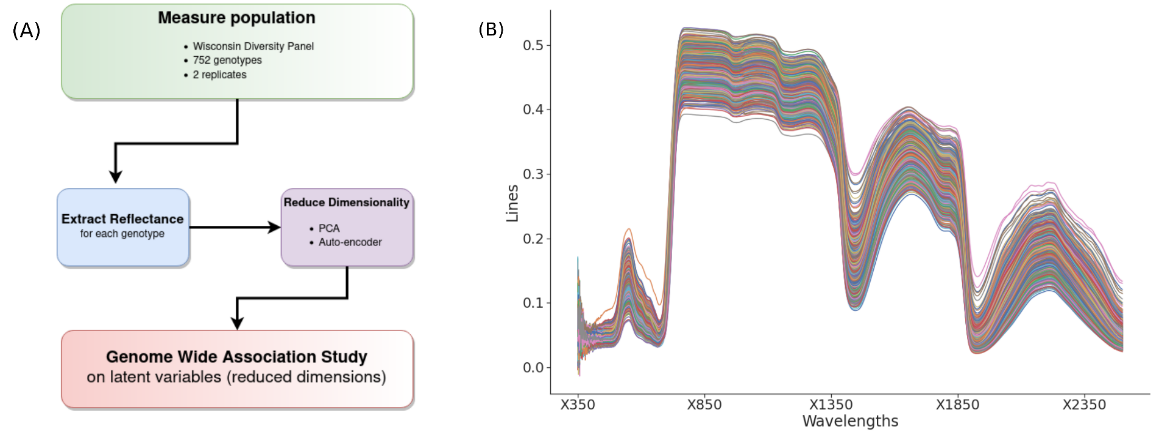

Figure 1. (A). Workflow for latent space phenotyping using maize leaf hyperspectral reflectance data. (B). Raw reflectance values at each wavelength for all maize plants. Each unique colour represents a unique genotype.

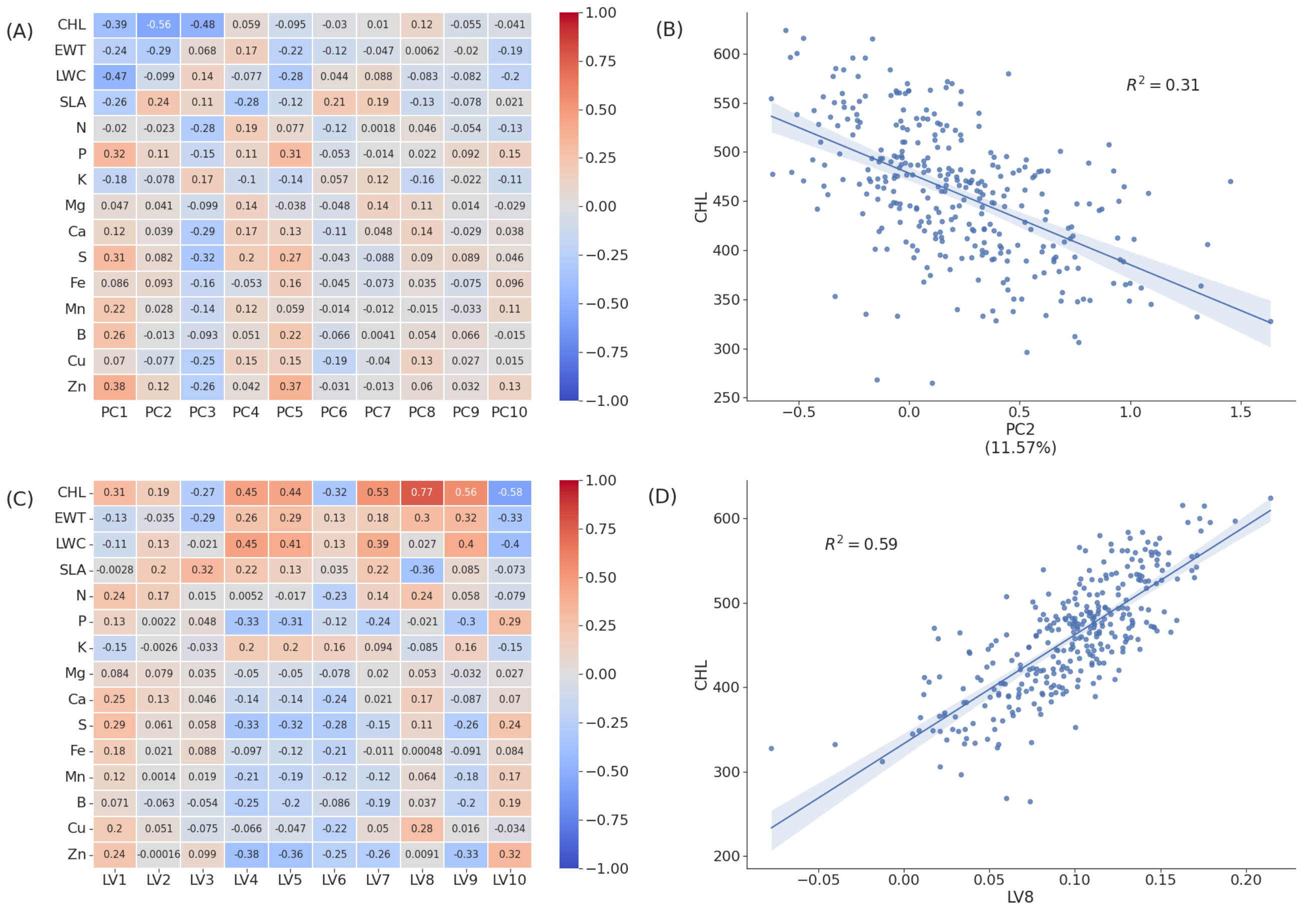

Figure 2. (A). Spearman correlation coefficients between individual principal components and ground truth measurements for 15 traits scored for 243-318 plots as part of this study. (B). Correlation between chlorophyll content and principal component 2 derived from the auto-encoder. (C). Spearman correlation coefficients between latent variables and molecular traits estimated in this study. (D). Correlation between chlorophyll content and latent variable 8.

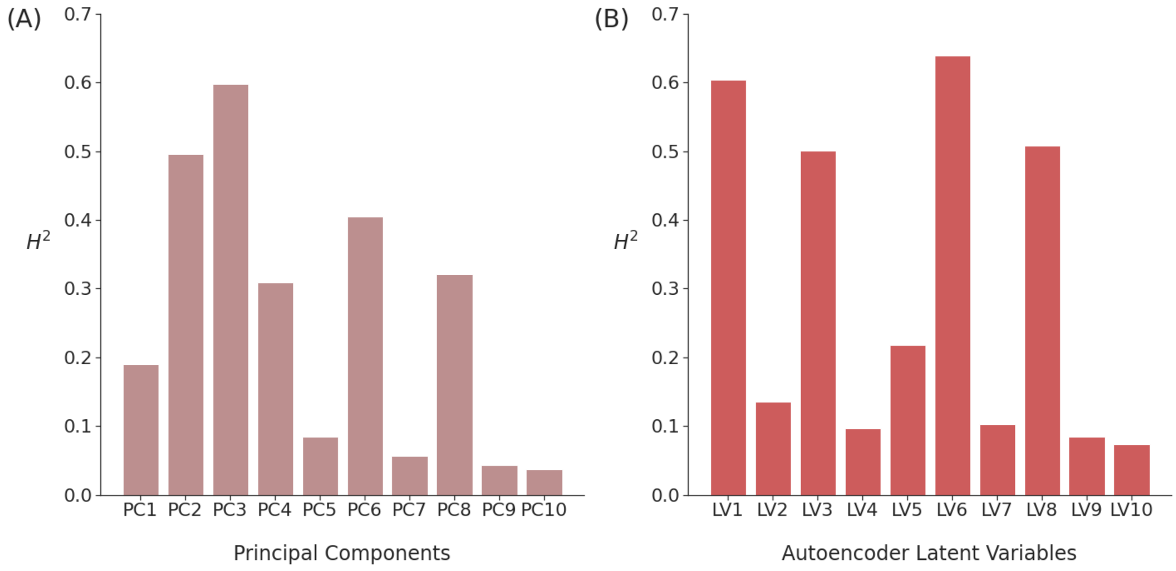

Figure 3. (A). Heritability (ie. The proportion of total variance explained by genetics) of each principal component (PC). (B). Heritability of each auto-encoder latent variable (LV).

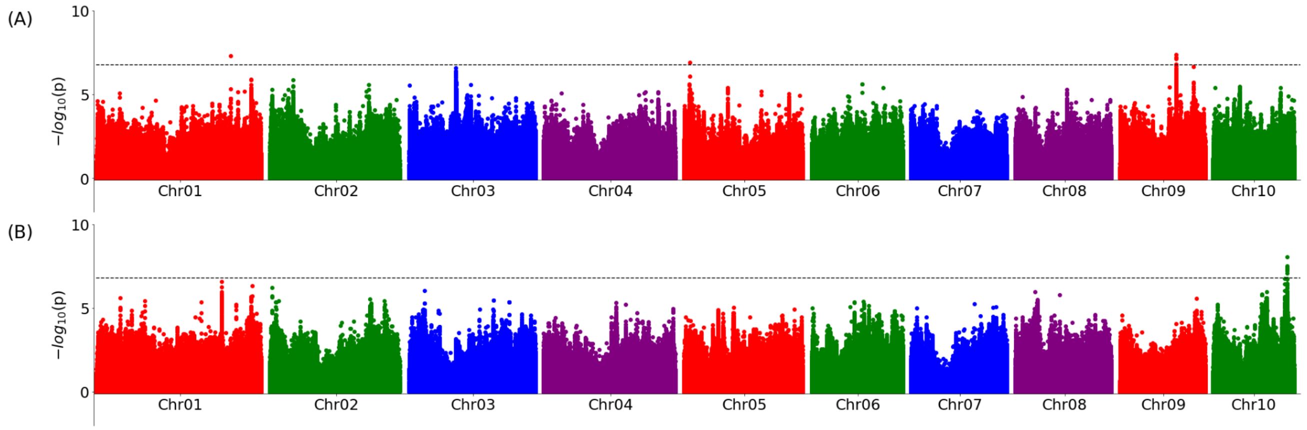

Figure 4. (A). Genetic markers significantly associated with latent variable 3 of the autoencoder trained on leaf reflectance data. (B). Genetic markers significantly associated with principal component 8 derived from leaf reflectance data. Each point indicates the physical position and statistical significance of an individual genetic marker. Dashed black lines indicate the threshold for statistical significance of 1.67 x 10^{-7} which was derived from a p-value cut off of 0.05 and a bonferroni adjustment for 299,685 effective snps.

References

Mazaheri, M., Heckwolf, M., Vaillancourt, B. *et al.* Genome-wide association analysis of stalk biomass and anatomical traits in maize. *BMC Plant Biol* **19,** 45 (2019).

Gage, J. L., Richards, E., Lepak, N., Kaczmar, N., Soman, C., Chowdhary, G., Gore, M. A., & Buckler, E. S. (2019). In‐field whole‐plant maize architecture characterized by subcanopy rovers and latent space phenotyping. *The Plant Phenome Journal*, *2*(1), 1–11.

Grzybowski, M., Wijewardane, N. K., Atefi, A., Ge, Y., & Schnable, J. C. (2021). Hyperspectral reflectance-based phenotyping for quantitative genetics in crops: Progress and challenges. *Plant Communications*, *2*(4), 100209.

Ubbens, J., Cieslak, M., Prusinkiewicz, P., Parkin, I., Ebersbach, J., & Stavness, I. (2020). Latent Space Phenotyping: Automatic Image-Based Phenotyping for Treatment Studies. *Plant Phenomics*, *2020*, 5801869.