

Using a Locally Connected Autoencoder to Identify Candidate Alleles in *Sorghum bicolor*

Mihael Cudic

Carnegie Mellon University, Robotics Institute

Abstract:

Motivation: By isolating Single Nucleotide Polymorphisms (SNPs) thought to be correlated to a phenotype, further research can be done to better target genomes to potentially eradicate diseases or alter certain properties in organisms.

Limitations: (1) Allele panels have extremely high dimensionality (2) Individuals cluster into sub-populations (3) Current approaches suffer from spurious false positives

Objectives: (1) Use an unsupervised approach to extract features in the allele panels (2) Find encoding dimensions correlated to phenotypes (3) Isolate candidate alleles for further investigation

Dataset: *Sorghum Bicolor* Allele Panels

- Allele panels contained 345 accessions with 232,303 SNPs for each accession [1]. Only 125,980 SNPs were used.

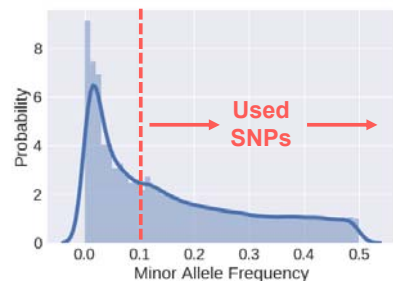


Figure 1: Minor allele frequencies for all 232,303 SNPs

- The *Sorghum bicolor* originated from 37 locations with 49.8% of individuals originating from only 5 locations

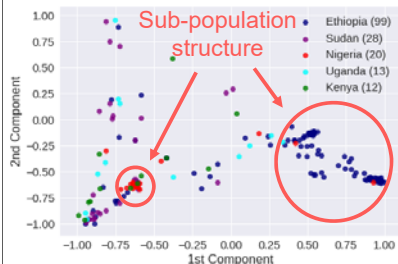


Figure 2: The 1st and 2nd normalized principal components for individuals originating from the 5 most common regions

Methodology:

High Dimensional Allele Panels

Feature Extraction

Link Features to Phenotype

Identify Candidate Alleles

Use a locally connected autoencoder (fig. 3) [2]

Find correlated encoding nodes

Backpropagate gradients to find average SNP contribution [3]

Locally Connected Autoencoder:

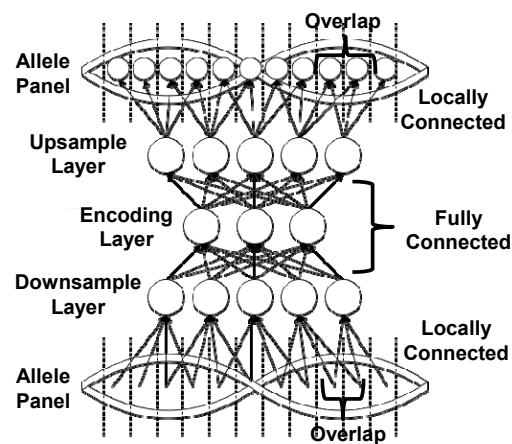


Figure 3: Summarized locally connected autoencoder topology used for feature extraction

Results:

- Correlations were found for fresh weight, stalk height, lignin percent and anthesis date
- S1_18096891, S2_2633689, S2_2793224, S6_7123769, and S6_61098274 were identified as candidate alleles for fresh weight and stalk height

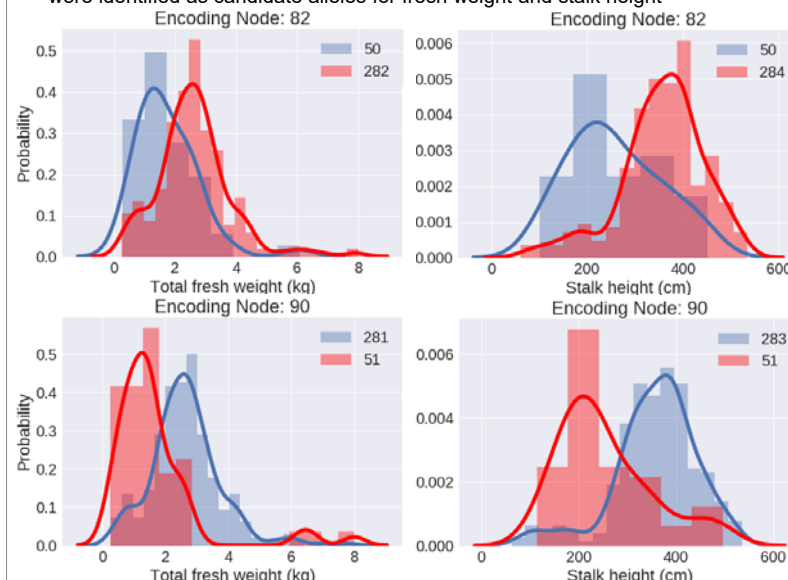


Figure 4: Clusters created by encoding node 82 and 90 on total fresh weight and stalk height data. The individuals in each cluster can be seen in the legend.

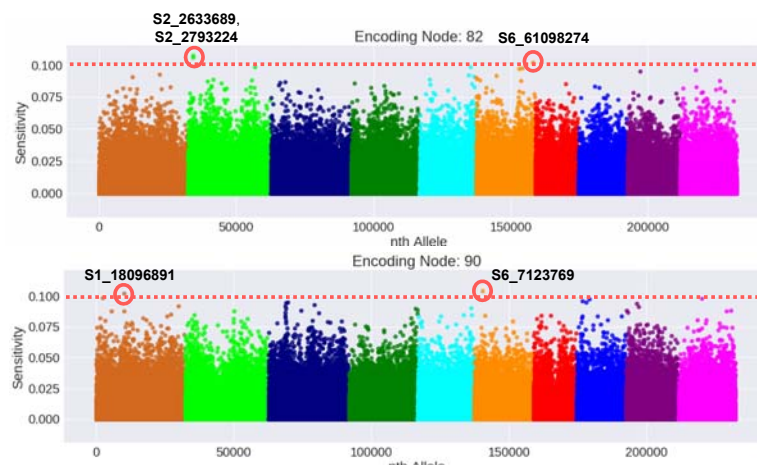


Figure 5: Manhattan plots of the SNPs average contribution to encoding node 82 and 90. Each color corresponds to a new chromosome.

Future Works:

- Test on other allele panels
- Better handle SNPs with highly skewed minor allele frequencies
- Validate candidate alleles through image data
- Confirm candidate alleles through additional biological research

Acknowledgements:

Thank you to Dr. Stephen Nuske for introducing me to a challenging problem in genetics as well as mentoring me during my time at CMU. I would like to additionally thank Dr. Nuske's group and all other groups working on the TERRA project for being supportive yet critical of my research. Finally, I am thankful for Dr. Dolan, Ms. Burcin and Ms. Maeda for supporting me throughout my stay at CMU.

References:

- Z. W. Brenton, "A genomic resource for the development, improvement, and exploitation of sorghum for bioenergy," *Genetics*, vol. 204, no. 1, pp. 21–33, 2016.
- Q. V. Le, "Building high-level features using large scale unsupervised learning," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 8595–8598.
- I. Dimopoulos, "Neural network models to study relationships between lead concentration in grasses and permanent urban descriptors in athens city (greece)," *Ecological modelling*, vol. 120, no. 2, pp. 157–165, 1999.