**Capstone 3 is an example of Unsupervised Learning Clustering**

**Goal**: Group or cluster soybean genotypes based on genetic similarity (Genetic diversity Analysis)

1. **Introduction/Background:**

Understanding the genetic structure of germplasm collections is a prerequisite for effective and efficient use of crop genetic resources. Hierarchical clustering using all genotype data has a long history and remains a popular way of representing genetic relations between individuals (Odong et al., 2011a). Here, I implemented clustering of USDA soybean germplasm collection based on genetic similarity (<https://soybase.org/snps>). This soybean Germplasm includes 18,480 domesticated and 1168 wild soybean accessions introduced from 89 countries or developed in the United States, consists of wild, landrace and North American cultivars (<https://www.ars.usda.gov/midwest-area/urbana-il/soybeanmaize-germplasm-pathology-and-genetics-research/>). This collection was genotyped with the SoySNP50K Bead Chip consists of >50K single-nucleotide polymorphisms (SNPs) (Song et al., 2015)

1. **Data source:**

2a). The SoySNP50K iSelect BeadChip has been used to genotype the USDA soybean germplasm (Song et al., 2015). Complete data set for 20,087 G. max and G. soja accessions genotyped with 42,509 SNPs is available for Wm82.a2 in vcf format <https://soybase.org/snps/soysnp50k_wm82.a2_41317.vcf.gz>.

2b). Minor Allele Frequencies (MAF) of SoySNP50K SNPs. Contains 40,841 of the 47,337 SNPs (86%) had minor allele frequencies >10% among the landraces, elite cultivars and the wild soybean accessions <https://soybase.org/snps/snp50k_maf.txt>

2c). Table S1, from Song et al., 2015. Fingerprinting soybean germplasm and its utility in genomic research. G3: Genes| Genomes| Genetics 50(10):1999-2006.). Table S1. Description of 19,648 Glycine max and G. soja accessions genotyped with the SoySNP50K BeadChip. Data include the PI (Plant Introduction) number, Species (G. max vs. G. soja), Country of origin, Cultivar name if applicable, Maturity Group, 99.9% similar to another accession (Y/N), Landrace or North American Elite cultivars used for analysis and seed weight of accessions used in the association analysis of seed weight.

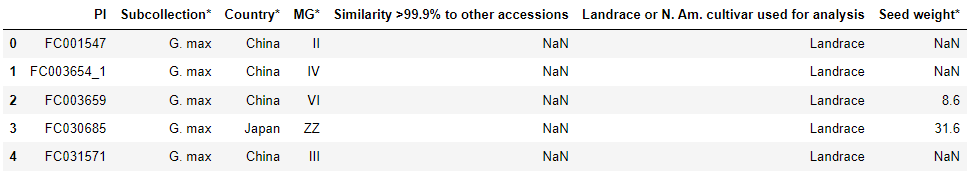
2d). List of 19649 soybean accessions tested in this study. [Soy\_USDA\_Accessions\_descritpion.xlsx](file:///C:\Users\gellima\Desktop\Springboard_DScience\CAPSTONE_3\'Soy_USDA_Lines_descritpion.xlsx')

1. **Data wrangling/Exploratory data analysis:**

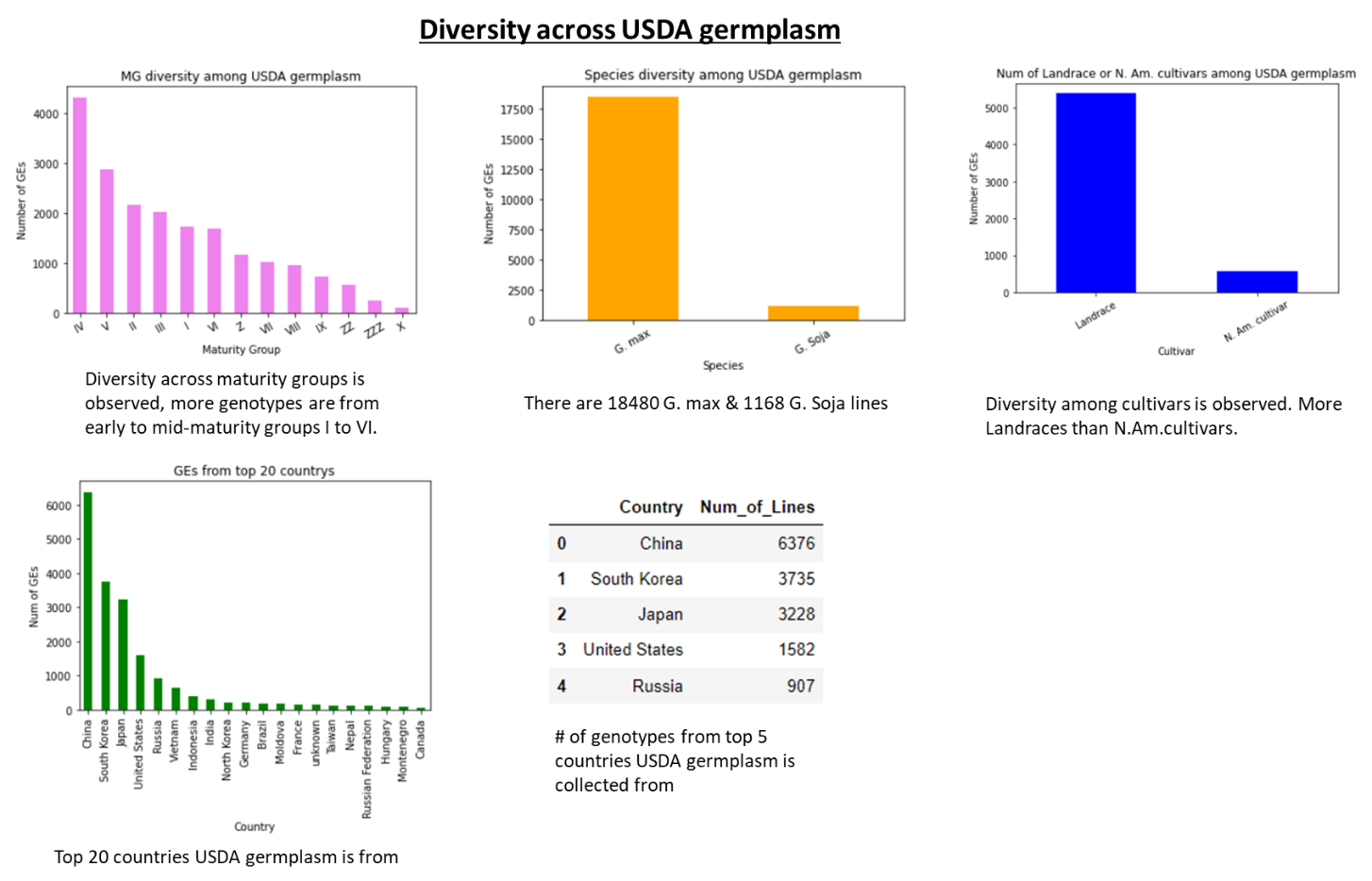
**3a) Soybean USDA accessions**

[Soy\_USDA\_Accessions\_descritpion.xlsx](file:///C:\Users\gellima\Desktop\Springboard_DScience\CAPSTONE_3\'Soy_USDA_Lines_descritpion.xlsx') contains

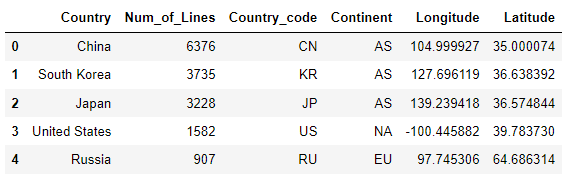
* This dataset consists of 19649 soybean accessions coming from 89 countries



* Observed the following diversity among USDA accessions

****

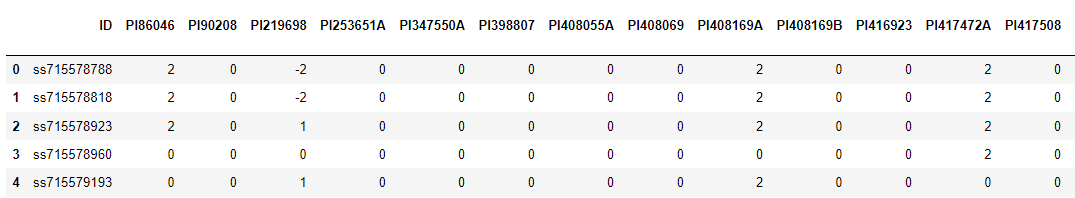
* Checked the geographic distribution of the USDA accessions; Country names are converted to alpha 2 country codes. For example, United States to US. Python’s pycountry-convert package is used to handle the conversion. Get longitude and latitude of continent using python’s geopy library. Plotted number of GE’s/country or continent on worldmap using python’s folium library.



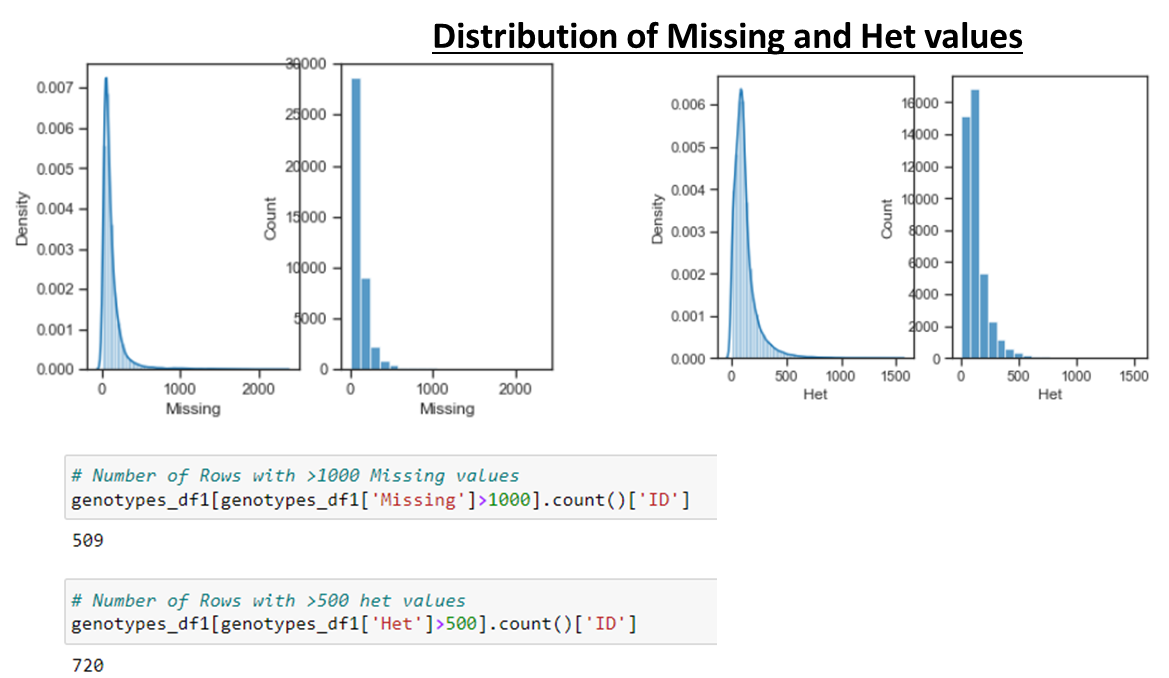
****

**3b). Soybean 50k SNP data**

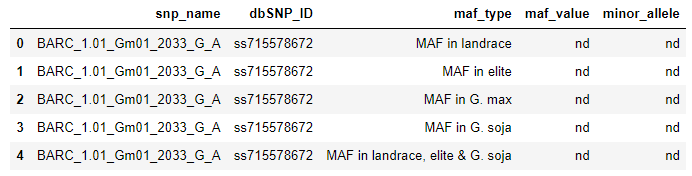
* vcf file contains data related to SNPs, chromosomes, positions, genotype scores for samples.
* allel.read\_vcf reads the vcf file as dictionary, since data type of each key values are different (1d, 2d, 3d arrays); converted all into 2D arrays for easy concatenation.
* scikit-allel has GenotypeArray() class, which added some convenient functionality to an array of genotype calls. Noticed that we have only 1 alternate allele, indicates SNPs are biallelic.
* Then, reshaped genotype array to view it as haplotypes by dropping the ploidy dimension using genotypes.to\_haplotypes() function. Then, created dataframe of haplotypes by transposing the array.
* Genotyping data calls: 0 reference allele, 2 alternate allele, 1 het, -2 consider as missing. Added sample names as column headers of the dataframe



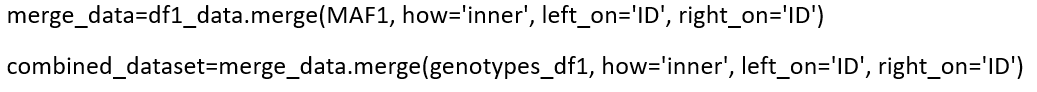
* Distribution of missing and het values indicates 509 entries having >1000 missing values and 720 entries with >500 het values.



**3c). Minor allele frequency dataset**



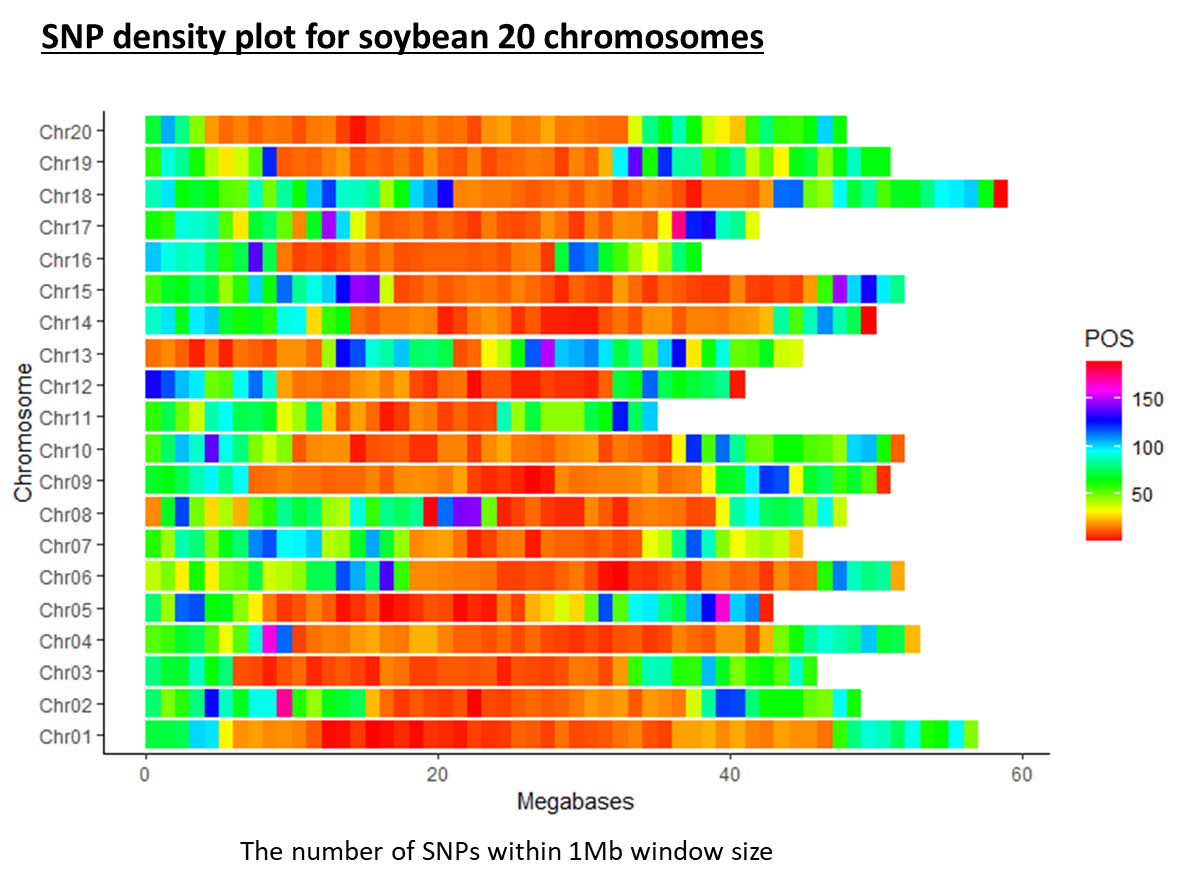
* Since maf\_type column contains 5 different variables, several duplicate values of snp\_name, dbSNP\_ID are observed. So, I changed the MAF to pivot table and then converted to dataframe (reduces the row count i.e 1 row per each snp\_name or ID)
* Concatenated genotyping data with MAF dataframe



1. **QC/filter combined dataset:**

**4a). Soy 50K data/SNP Filtering:** There are 12 SNPs having >10% missing. But no SNPs having >10% het calls. Later het calls for the remaining loci were changed as missing data. Then, dropped 37 monomophic SNPs (either Ref\_allele or Alt\_allele count 0). Followed by, dropped SNPs having MAF<0.05.

* + - Finally, 36729 SNPs remained after dropping SNPs with >10% missing data, monomorphic SNPS, MAF<0.05
    - Distribution of these filtered 36.7k SNPs across 1 mega base region of 20 chromosomes of soybean is shown below.



**4b). Soybean USDA accessions filtering:** There are 691 samples having >10% missing (138 samples with 15% missing, 20 samples having >20% missing data). Since there aren't many samples with higher missing data, I dropped samples with >10% missing data. After dropping samples with >10% missing data, 19396 samples are left.

* Based on the pair-wise genetic similarity of the accessions calculated from the 42,509 SNPs, 4299 G. max (23%) and 362 G. soja (31%) accessions having at least >99.9% identical to another accession in the collection

|  |  |  |
| --- | --- | --- |
| **Table. Number of accessions in the USDA Soybean Germplasm Collection with similarity >99% based on SNP comparisons** | | |
| **Similarity Among Accessions** | **G.Soja** | **G.max** |
| Accessions >99% similar to another accession | 362 | 4299 |
| Proportion of accessions with >99% similar (% | (362/1168) =31% | (4299/18480) =23.26 |

* A total of 18480 G. max & 1168 G. Soja accessions were used for further analysis.

# Machine learning models implemented

# Since this dataset doesn’t have labels, I implemented Unsupervised Learning (USL). USL is a class of machine learning technique used to find patterns in data instead of predicting target variable like in supervised learning.

# 5a) Cluster analysis (KMeans)

# Grouping/clustering the genotypes/accessions based on similarity/patterns. I implemented KMeans clustering model from python’s sklearn library, partitions dataset into k clusters based on similarity in which each observation belongs to the cluster with nearest mean.

# Converted dataframe to 2D numpy array before implementing KMeans algorithm.

# Plot of inertia values of clustering with different K values showed that K means model with 5 clusters had relatively low inertia, so it looks like 5 clusters would be a good choice for this dataset.

# 

# Then, grouped accessions into 5 clusters and predicted the label/cluster for each line.

# 

# Then, evaluated the clustering with cross tabulation. The cross tabulation shows that the samples separate really well into 5 clusters. But depending on the type of data you work with, the clustering may not always be this good.

# Is there anything you can do in such situations to improve your clustering?

# YES, to give every feature a chance, data needs to be transformed. So, all features will have equal variance during preprocessing. For example: StandardScaler: standardizes features by removing the mean and scaling to unit variance i.e transforms each feature to have mean 0, variance 1 and standardized features are very informative. Normalizer: rescales each sample independently of the other

# In this dataset, I implemented clustering with KMeans and also used transformations like StandardScaler and Normalizer (n\_clusters=5).

# 

# The cluster labels generated with Kmeans, with and without StandardScaler/Normalizer, are different. However, same GENames were grouped together though the cluster label is different with transformations.

# 

# 5b) Visualizing hierarchies (hierarchial clustering/Dendrogram)

# Hierarchial clustering arranges samples into hierarchy of clusters

# 

# 

# SciPy hierarchical clustering doesn't fit into a sklearn pipeline, so I used normalize() function from sklearn.preprocessing instead of Normalizer

# 

# 

# Since there are 19,400 GENames/accessions to cluster, the hierarchical clustering seems clumsy. So, I exported the clusters and grouped them based on Species, Cultivar and Relative Maturity groups.

# Species: All clusters except 3, had higher G.max accessions. Cluster 3 had G.soja accessions.

# Cultivar: All clusters had mostly Landraces.

# Relative maturity:

# Cluster 0: Many mid-maturity group (IV, V, VI)

# Cluster 1: Many early maturity group accessions (I, II)

# Cluster 2: Spread across mid to early-late maturity groups (IV to IX)

# Cluster 3: Group IV, V

# Cluster 4: Spread across maturity groups.

# 

# Conclusions and future work:

# The United States Department of Agriculture (USDA), Soybean Germplasm Collection includes 18,480 domesticated soybean and 1168 wild soybean accessions introduced from 84 countries or developed in the United States. This collection was genotyped with the SoySNP50K BeadChip containing greater than 50K single-nucleotide polymorphisms.

* + Any polymorphic SNP with a rate of missing and heterozygous allele calls greater than 0.1 among the accessions was eliminated. The heterozygous allele calls in the remaining loci were set as missing in the subsequent analysis. Filtering of 50K SNPs in this dataset showed well distribution of 36.7k SNPs across 1 mega base region of 20 chromosomes A total of 18480 G. max & 1168 G. Soja accessions were used for further analysis.

# Since this dataset doesn’t have labels, I implemented Unsupervised Learning (USL). Implemented KMeans clustering algorithm.

# Plot of inertia values of clustering with different K values showed that K means model with 5 clusters had relatively low inertia, so it looks like 5 clusters would be a good choice for this dataset.

* In this dataset, I implemented clustering with KMeans and also used transformations like StandardScaler and Normalizer (n\_clusters=5).The cluster labels generated with Kmeans, with and without StandardScaler/Normalizer, are different. However, same GENames were grouped together though the cluster label is different with transformations.

# Also implemented hierarchical clustering along with and without Normalizer. Since there are 19,400 GENames/accessions to cluster, the hierarchical clustering seems clumsy. So, I exported the clusters and grouped them based on Species, Cultivar and Relative Maturity groups.

# For future work, will perform clustering with Principal Component Analysis (PCA).

# Validate the cluster groups found in this study by applying the KMeans, Hierarchial models to the established germplasm from NA breeding programs and evaluating the genetic similarity of the accessions used in this study to the products/varieties released using these accessions as parents.

# Also evaluating the association of genetic similarity with traits/phenotypes measured on these accessions would give more value to these accessions for breeding programs.

# Besides, research showed that population structure is more efficiently captured by PCA-reduced data (Patterson et al., 2006) while improving hierarchical clustering (van Heerwaarden et al., 2013). There is a growing interest in PCA-based clustering (Lee et al., 2009; Paschou et al., 2007; van Heerwaarden et al., 2010) calls for an understanding of the potential benefits of PCA-based data reduction for improving hierarchical clustering. So, I am planning to evaluate if PCA improves clustering of the accessions.

# References:

# Odong, T.L., van Heerwaarden, J., Jansen, J., van Hintum, Th.J.L., and van Eeuwijk, F.A.. 2011a. Determination of genetic structure of germplasm collections: Are traditional hierarchical clustering methods appropriate for molecular marker data? Theor. Appl. Genet. 123:195–205.

# Song, Qijian, David L. Hyten, Gaofeng Jia, Charles V. Quigley, Edward W. Fickus, Randall L. Nelson, and Perry B. Cregan. 2015. Fingerprinting soybean germplasm and its utility in genomic research. G3: Genes| Genomes| Genetics 50(10):1999-2006.

# van Heerwaarden, J., Odong, T.L., and van Eeuwijk, F.A.. 2013. Maximizing genetic differentiation in core collections by PCA-based clustering of molecular marker data. Theor. Appl. Genet. 126(3):763–772.

# Lee, C., Abdool, A., and Huang, C.H.. 2009. PCA-based population structure inference with generic clustering algorithms. BMC Bioinf. 10(S1):S73.

# Paschou, P., Ziv, E., Burchard, E.G., Choudhry, S., Rodriguez-Cintron, W., Mahoney, M.W., and Drineas, P.. 2007. PCA-correlated SNPs for structure identification in worldwide human populations. PLoS Genet. 3:1672–1686.

# van Heerwaarden, J., Ross-Ibarra, J., Doebley, J., Glaubitz, J.C., Jesus Sanchez, D.E., Gonzalez, J., Gaut, B.S., and Eguiarte, L.E.. 2010. Fine scale genetic structure in the wild ancestor of maize (Zea mays ssp. parviglumis). Mol. Ecol. 19:1162–1173.