ISSN 1479-2621
Plant Genetic Resources: Characterization and Utilization (2018) **16**(6); 513–523
doi:10.1017/S1479262118000102

# Population and quantitative genomic properties of the USDA soybean germplasm collection

Alencar Xavier[1,2], Rima Thapa[1], William M. Muir[3] and Katy Martin Rainey[1]*

[1]*Department of Agronomy, Purdue University, West Lafayette, IN 47907, USA,* [2]*Quantitative Genetics, Dow AgroSciences, Indianapolis, IN 46268, USA and* [3]*Department of Animal Sciences, Purdue University, West Lafayette, IN 47907, USA*

**Received 7 November 2017; Accepted 23 March 2018 – First published online 23 April 2018**

## Abstract

This study is the first assessment of the entire soybean [*Glycine max* (L.) Merr] collection of the United State Department of Agriculture National Plant Germplasm System (USDA) reporting quantitative and population genomic parameters. It also provides a new insight into soybean germplasm structure. Germplasm studies enable plant breeders to incorporate novel genetic resources into breeding pipelines to improve valuable agronomic traits. We conducted comprehensive analyses on the 19,652 soybean accessions in the USDA-ARS germplasm collection, genotyped with the SoySNP50 K iSelect BeadChip SNP array, to elucidate the quantitative properties of existing subpopulations inferred through hierarchical clustering performed with Ward's *D* agglomeration method and Nei's standard genetic distance. We found the effective population size to be approximately 106 individuals based on the linkage disequilibrium of unlinked loci. The cladogram indicated the existence of eight major clusters. Each cluster displays particular properties with regard to major quantitative traits. Among those, cluster 3 represents the tropical and semi-tropical genetic material, cluster 5 displays large seeds and may represent food-grade germplasm, and cluster 7 represents the undomesticated material in the germplasm collection. The average $F_{ST}$ among clusters was 0.22 and a total of 914 SNPs were exclusive to specific clusters. Our classification and characterization of the germplasm collection into major clusters provides valuable information about the genetic resources available to soybean breeders and researchers.

**Keywords:** fixation index, genetic correlation, phylogenomics, structure, unsupervised machine learning

## Introduction

Soybean [*Glycine max* (L.) Merr.] is the second most important crop worldwide used for animal feed and human foods. Soybean is cultivated globally, especially in the USA, Brazil, Argentina and China. According to Hymowitz (2008), Southeast Asia is the geographical origin of the genus *Glycine*. The closest extant wild relative of domesticated soybean is *Glycine soja*, which is considered to be its undomesticated progenitor over 3000 years ago (Doebley *et al*., 2006). *G. max* and *G. soja* possess valuable morphological, physiological and biochemical traits (Carter *et al*., 2004a, b; Chan *et al*., 2012) which can be beneficial for crop improvement. Hybrids between accessions of the two species can successfully produce viable and fertile progeny (Singh and Hymowitz, 1989; Carter *et al*., 2004a, b; Kuroda *et al*., 2013).

Many studies have reported metrics of genetic diversity, structure and quantitative trends among complex traits in soybean and its wild progenitor; however, these have limited genetic resources and marker data that were available

*Corresponding author. E-mail: krainey@purdue.edu

(Brown-Guedira *et al.,* 2000; Guo *et al.,* 2010). Previous studies have been performed upon soybean germplasm attempting to provide a better understanding of the soybean genome footprints (Song *et al.,* 2015; Zhou *et al.*, 2015). Among these footprints, some investigations focused on the bottlenecks associated with domestications (Hyten *et al.,* 2006; Zhao *et al.,* 2015), and others on the diversity and possible applications for breeding (Bandillo *et al.,* 2015; Jarquin *et al.,* 2016). Yet, there exists a gap the regards the simultaneous analysis of quantitative and population genomic parameters. The United States Department of Agriculture (USDA) Soybean germplasm collection was officially established in 1949 and served as an important resource for soybean researchers worldwide (Carter *et al.,* 2004b). According to a report from Carter *et al.* (2004b), the USDA soybean germplasm collection consisted of five sub-collections, including *G. max*, *G. soja*, perennial *Glycine* species, genetic stocks and domestic cultivars.

The genetic variation confined within the USDA germplasm collection relates to important traits for soybean improvement (Bandillo *et al.,* 2015). Genome-wide associations on seed protein and oil conducted in approximately 12,000 soybean accessions were able to identify genomic regions with strong signals in chromosomes 15 and 20 (Bandillo *et al.,* 2015) and population structure analysis on approximately 14,000 accessions reported the existence of subpopulations within the collection (Bandillo *et al.,* 2015).

This study combines the entire USDA soybean germplasm collection with high-density marker data and passport data to greatly increase our understanding of the population and quantitative genomic properties of the domesticated germplasm and its wild relatives. To create a profile of the collection, we identified and characterized subpopulations, and estimated the heritability and genetic correlation among major agronomic traits at the whole-germplasm level. We report phylogenomic analyses for the entire USDA germplasm collection, comprising 19,652 accessions of *G. max* and *G. soja* genotyped with 42,509 SNP markers. Hierarchical cluster analysis showed that the USDA collection is divided into eight well-defined clusters inferred in a non-subjective manner. Our results will assist soybean breeders with a selection of accessions to use for genetic improvement.

## Methods

### Germplasm collections

We studied the 19,652 accessions from the USDA Soybean Germplasm Collection. Previous efforts to describe this germplasm are provided by Bandillo *et al.* (2015) and Jarquin *et al.* (2016). Detailed information on any of the accessions is available through the USDA Germplasm Resources Information Network (GRIN) database (www.ars-grin.gov).

The USDA germplasm collection consists of various countries around the world, especially from East Asia (Song *et al.,* 2015). Some of the accessions in the USDA collection have resistance against various diseases and pests (Song *et al.,* 2015; Chang *et al.,* 2016) and also have unique genes for composition modification and yield (Zhang *et al.,* 2016). In spite of the overall benefits from genetic diversity, more research is needed to identify sources of useful alleles for crop improvement (Chan *et al.,* 2012).

### Genotypic data

Song *et al.* (2013) generated the genotypic data used for these analyses by genotyping each accession in the USDA Soybean Germplasm Collection with 42,509 SNPs arrayed on a SOYSNP50 K iSelect BeadChip. We accessed the data for all accessions in the *G. soja* and *G. max* collections from the SoyBase website (Grant *et al.,* 2009) on 5 December 2014. We coded allelic genotype {AA, Aa, aa} for the imputation procedure as {0, 1, 2} (Stranden and Christensen, 2011), where AA is homozygous towards the reference genome Williams82. Then we imputed missing loci using random forest regression (Stekhoven and Buhlmann, 2012), one chromosome at a time.

### Distance matrix and phylogenetic tree

There are many metrics for dissimilarity among genotypes (Reif *et al.,* 2005); we assessed several combinations of linkage and dissimilarity (James *et al.*, 2013) unsupervised machine-learning strategy (Xavier *et al.,* 2016) to obtain the clearest distinction of clades. The bi-plot projection of genetic dissimilarity the germplasm was based on multidimensional scaling (Cox and Cox, 2000), as implemented in the R function *cmdscale*.

The dissimilarities assessed included Nei's genetic distance, Edwards' genetic distance, Euclidean distance, Manhattan distance, and Canberra distance, the genetic distances (Nei's and Edwards') were utilized as implemented in the R package adagenet (Jombart and Ahmed, 2011) and the other methods (Euclidean, Manhattan and Canberra) are built-in from the R function *dist*.

The linkage methods evaluated for each dissimilarity to build phylogenetic trees were: Neighbour joining, weighted and unweighted pair group method with arithmetic mean (UPGMA and WPGMA), Ward $D_1$ and War $D_2$. Neighbour joining was used as implemented in the R package APE (Paradis *et al.,* 2004), the other methods are available in the built-in R function *clust*.

The phylogenetic tree from each combination of linkage and dissimilarity is available in the Supplementary material

https://doi.org/10.1017/S1479262118000102 Published online by Cambridge University Press

(File S1). Due to the best distinction among clades, we proceeded with the analyses using Nei's standard genetic distance (Nei, 1972) as the measure of dissimilarity, and Ward's $D_1$ as the linkage or clustering method (Murtagh and Legendre, 2014). Nei's distance between two genotypes is estimated from molecular genotypes as

$$d_{(i,j)} = \frac{-\ln\left(\boldsymbol{x}_i'\boldsymbol{x}_j\right)}{\sqrt{\boldsymbol{x}_i'\boldsymbol{x}_j}}, \quad (1)$$

where $\boldsymbol{x}$ represents a vector comprising the genotypic information and alleles {*AA*, *Aa*, *aa*} are coded as {[2, 0] [1, 1] [0, 2]}. To compute the genetic distance among populations, we used the Nei's standard genetic distance.

Agglomeration through Ward's $D$ endeavours to cluster genotypes by maximizing between-group variance (Murtagh and Legendre, 2014). Thus, once it increases the changes of grouping genotypes with markers fixed for the same allele, it has interesting grouping properties from the genetic perspective; consequently, it favours a more accurate identification of subpopulations (Wright, 1965). Ward's $D_1$ may display changes in topology when employed over Euclidean distances (Reif *et al.*, 2005). The clustering was performed by cutting the phylogenetic tree at a height of 100, where clades presented clear distinction (Fig. 1a).

## Passport information

The passport information describing accession origins was provided by Dr Randall Nelson, research geneticist at USDA-ARS, and is available in the Supplementary material (File S2). Traits used for this analysis were domestication status (MAX), indicating whether an accession is *G. max* (MAX = 1) or *G. soja* (MAX = 0); weight of 100 seeds (SW), maturity group (MG), grain yield in grams per plant (YLD), plant height in centimetres (HT), lodging score 1–5 (LDG), percentage of seed oil content (OIL) and percentage of seed protein content (PRO).

## Quantitative analysis of complex traits

We obtained heritability and genetic correlations from the covariance components estimated using the multivariate mixed linear model implemented in GISBBS3F90 (Misztal *et al.*, 2002). The multivariate model fits all traits at once, while the linear model for each trait can be described as

$$\boldsymbol{y}_k = 1\boldsymbol{\mu}_k + \mathbf{Z}_k\mathbf{u}_k + \mathbf{e}_k. \quad (2)$$

where for the $k$th trait, $\mathbf{y}_k$ is the vector of phenotypes, $\boldsymbol{\mu}_k$ is the intercept, $\mathbf{Z}_k\mathbf{u}_k$ corresponds to the genetic term treated as a random effect, provided that $\mathbf{Z}_k$ is the incidence matrix of genotypes, $\mathbf{u}_k$ represents the regression coefficients known as breeding values, and $\mathbf{e}_k$ is the vector of residuals. The variance of the multivariate model is described as

$$\mathrm{Var}(\mathbf{Y}) = \left(\mathbf{A} \otimes \boldsymbol{\Sigma}_a + \mathbf{I} \otimes \boldsymbol{\Sigma}_e\right). \quad (3)$$

where $\mathbf{A}$ is the genomic relationship matrix estimated from the molecular markers, $\boldsymbol{\Sigma}_a$ is the genetic covariance component matrix that comprises genetic variances $\left(\sigma^2_{a_i}\right)$ in the main diagonal and genetic covariances in the off-diagonal ($\sigma_{a_{ij}}$), $\mathbf{I}$ is an identity matrix, and $\boldsymbol{\Sigma}_e$ is the residual covariance matrix with residual variances ($\sigma^2_{e_i}$) in the main diagonal and residual covariances in the off-diagonal ($\sigma_{e_{ij}}$). Using the elements from $\boldsymbol{\Sigma}_a$, we estimated the genetic correlations as

$$\rho_{a_{ij}} = \frac{\sigma_{a_{ij}}}{\sigma_{a_i}\sigma_{a_j}}. \quad (4)$$

## Subpopulation differentiation analysis

The genomic structure of the broad germplasm panel provides insight into the demography, domestication and adaptation of soybean (Holsinger and Weir, 2009). It also identifies genomic regions which may possess little variation, or alternately, can be sources of exotic alleles. We used the $F_{ST}$ fixation index (Wright, 1949) to assess

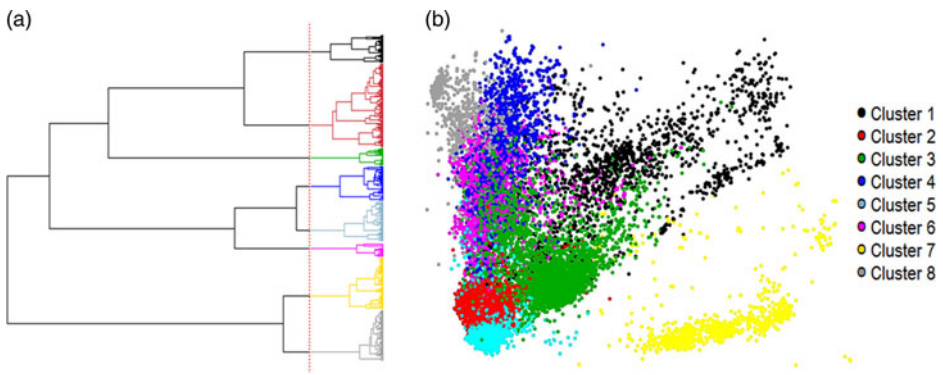


**Fig. 1.** Clustering of USDA soybean germplasm collection based on Nei's standard genetic distance clustered with Ward's $D_1$. (a) Cladogram and (b) Multidimensional Scale.

https://doi.org/10.1017/S1479262118000102 Published online by Cambridge University Press

differentiation among clusters, with the unbiased estimator of Weir and Cockerham (1984) implemented by Xavier *et al.* (2015), using the smoothing kernel strategy (Flori *et al.,* 2009) to discriminate between signature selection and domestication from drift.

### *Linkage disequilibrium (LD)*

Our analysis computed LD in terms of $r^2$ computed from pairs of SNPs, phasing molecular markers using the Expectation–Maximization algorithm (Slatkin and Excoffier, 1996). The computation was performed by the function *LD* implemented in the R package NAM (Xavier *et al.,* 2015).

The estimated effective population size was calculated using an approximation (Sved *et al.,* 2013; Waples *et al.,* 2014) based on the mean elements from the off-diagonal LD matrix ($\hat{r}^2$), as

$$\hat{N}_e \approx \frac{1}{3\hat{r}^2}. \quad (5)$$

## Results

### *Clusters*

Figure 1 presents the cladogram used to generate the eight clusters and the multidimensional scaling (MDS) representation obtained by best-fitting the *k*-dimensional representation (Mardia, 1978). The cladogram produced through Ward's $D_1$ agglomeration yielded clear clusters, although it may not be obvious that all lines are included in Fig. 1a. The MDS is directly comparable with the principal components biplot presented by Bandillo *et al.* (2015), who used only a subset of this dataset. Our clustering analysis contrast with Jarquin *et al.* (2016), where nine subpopulations were inferred from centroid agglomeration from the principal components.

A representation of clusters in terms of the countries that made major contributions to the germplasm collection, along with the population-wise genetic distance, is provided in the Supplementary material (File S1).

### *Quantitative trends*

The subpopulation inferred here present distinct genetic and phenotypic characteristics. The number of individuals in each cluster, alongside the average value for each phenotype under evaluation, is summarized in Table 1. This information provides an insight into the genetic resources available in the collection for breeding applications, also highlighting phenotypic discrepancies among clusters.

Most valuable traits, such as yield and seed composition, are complex due to their quantitative nature, meaning that they are controlled by many genes that each have small effects and that they are responsive to environmental stimuli. Table 2 shows the genetic associations among the agronomic traits. Quantitative genetic parameters contribute to the strategic use of novel germplasm in breeding programmes; for instance, programmes looking to increased seed weight could consider exploiting cluster 5.

Genetic parameters obtained from covariance components are measurements of the joint response of multiple traits to selection and drift, which are reflected on the heritabilities and genetic correlations (Xavier *et al.,* 2017). The very nature of this interaction can be associated to the phenotypic plasticity associated with fitness, physiological constraints of each phenotype, and the pleiotropy due to shared pathways the affect multiple traits simultaneously (DeJong and VanNoordwijk, 1992; Zera and Harshman, 2001). Despite the possibility of subpopulation to present slightly different quantitative genetic properties, Table 2 presents values estimated from the germplasm collection as a whole.

**Table 1.** Summary of cluster features

| Cluster | Freq | MAX (%) | SW | MG | YLD | HT | LDG | OIL | PRO |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 1619 | 99 | 11.3 | 3.8 | 1.9 | 94.9 | 3.5 | 17.3 | 43.8 |
| 2 | 3168 | 100 | 16.8 | 4.5 | 1.8 | 68.8 | 2.8 | 17.3 | 45.3 |
| 3 | 5066 | 100 | 12.5 | 5.6 | 1.6 | 107.4 | 3.3 | 16.6 | 45.8 |
| 4 | 2407 | 100 | 16.3 | 2.1 | 2.6 | 84.2 | 2.5 | 19.9 | 42.0 |
| 5 | 3087 | 100 | 19.2 | 4.5 | 1.7 | 71.8 | 2.4 | 18.1 | 43.9 |
| 6 | 2110 | 100 | 15.6 | 1.8 | 2.5 | 86.3 | 2.8 | 18.8 | 43.3 |
| 7 | 1208 | 2.7 | 1.6 | 4.6 | 0.5 | 77.4 | 4.9 | 11.2 | 47.0 |
| 8 | 987 | 99 | 15.9 | 3.2 | 3.2 | 92.9 | 1.9 | 20.0 | 42.1 |

Number of individuals (Freq), Percentage of domesticated soybeans (MAX), Weight of 100 seeds (SW), Maturity group (MG), Grain yield in grams per plant (YLD), Plant height in centimetres (HT), Lodging score 1–5 (LDG), Percentage of seed oil content (OIL), Percentage of seed protein content (PRO).

https://doi.org/10.1017/S1479262118000102 Published online by Cambridge University Press

**Table 2.** Genetic correlations (off-diagonal) and heritability (main diagonal, bold) inferred from the USDA soybean germplasm collection.

| | MAX | SW | FLO | MAT | OIL | PRO | HT | MG | YLD | LDG |
|---|---|---|---|---|---|---|---|---|---|---|
| MAX | **0.99** | −0.03 | −0.09 | −0.01 | −0.04 | 0.08 | 0.12 | 0.03 | 0.13 | 0.43 |
| SW | – | **0.62** | 0.07 | 0.06 | 0.22 | −0.09 | 0.02 | −0.19 | 0.46 | −0.11 |
| FLO | – | – | **0.34** | 0.85 | −0.29 | 0.03 | 0.31 | −0.07 | −0.02 | 0.13 |
| MAT | – | – | – | **0.38** | −0.25 | 0.06 | 0.38 | 0.36 | −0.17 | 0.21 |
| OIL | – | – | – | – | **0.65** | −0.83 | −0.04 | −0.07 | 0.36 | 0.01 |
| PRO | – | – | – | – | – | **0.62** | 0.06 | 0.15 | −0.24 | −0.02 |
| HT | – | – | – | – | – | – | **0.70** | 0.32 | 0.12 | 0.72 |
| MG | – | – | – | – | – | – | – | **0.82** | −0.32 | 0.14 |
| YLD | – | – | – | – | – | – | – | – | **0.52** | −0.01 |
| LDG | – | – | – | – | – | – | – | – | – | **0.56** |

Percentage of domesticated soybeans (MAX), Weight of 100 seeds (SW), Flowering date (FLO), Maturity date (MAT), Percentage of seed oil content (OIL), Percentage of seed protein content (PRO), Plant height in centimetres (HT), Maturity group (MG), Grain yield in grams per plant (YLD), Lodging score 1–5 (LDG).

## *Eight genetic groups*

Cluster analysis supports the existence of eight germplasm groups (Fig. 1). These clusters are likely to share specific phenotypic properties that characterize their breeding application or adaptation zone (Yamamichi and Innan, 2012). Some of the specific properties observed in each cluster are described below.

- **Cluster 1:** Displays the lowest seed weight and highest values of lodging among clusters of domesticated soybeans.
- **Cluster 2:** Mostly composed of Korean germplasm, presenting the lowest average plant height.
- **Cluster 3:** The largest group in the germplasm collection, it comprises the tropical and semi-tropical material. Highest average maturity group and plant height, also the lowest oil content among clusters of domesticated soybeans.
- **Cluster 4:** Mostly composed of germplasm from China and Russia. This cluster displays the lowest average seed protein content.
- **Cluster 5:** Displays the highest seed weight, which can be possibly associated with the food-grade soybean.
- **Cluster 6:** Includes Russian, North American and European material, with an average maturity group close to II.
- **Cluster 7:** Cluster of undomesticated germplasm, presenting the highest lodging scores and smallest seed weight.
- **Cluster 8:** Comprises US germplasm. It represents the highest yielding material and highest average for seed oil content, although the yield assessment was performed in US territory.

As we were able to find eight major clusters with shared features, it is also possible that each cluster holds subclusters. Previously, Shi *et al.* (2010), clustered the food-grade soybeans (corresponding to our cluster 5) and concluded that this subpopulation could be further stratified into clusters with distinct types of food-grade soybeans. Another example involves the undomesticated material, the representation of germplasm by country presented in the Supplementary material (File S1) indicates little overlap of wild relatives (cluster 7) across countries, which supports the existence of subpopulations (Wen *et al.,* 2009; Li *et al.,* 2010, 2013).

## *Signatures of genomic differentiation*

The average $F_{ST}$ among the eight clusters was 0.22 and the 95% quantile was 0.43, which is in agreement with results from Zhao *et al.* (2015). Figure 2 shows the $F_{ST}$ index with values presented in terms of $-\log(1-F_{ST})$. It is not obvious when the fixation index should be considered to be associated with a selection event and, therefore, $F_{ST}$ values are often compared within the SNP panel (Akey *et al.,* 2002). In this study, we express the fixation index in terms of $-\log(1-F_{ST})$ to provide a clearer representation of the genomic regions affected by selection-driven fixation, such that $F_{ST} = 0.9$ (i.e. 99% quantile) would represent 1 in this scale.

Several peaks across the genome indicate alleles close to within-cluster fixation (Wahlund effect), which are associated with the selection, genetic drift or population bottlenecks. The smoothing kernel (in blue) defines regions that characterize a selection event (Flori *et al.,* 2009), thus distinguishing these peak from genetic drift. The results are

https://doi.org/10.1017/S1479262118000102 Published online by Cambridge University Press

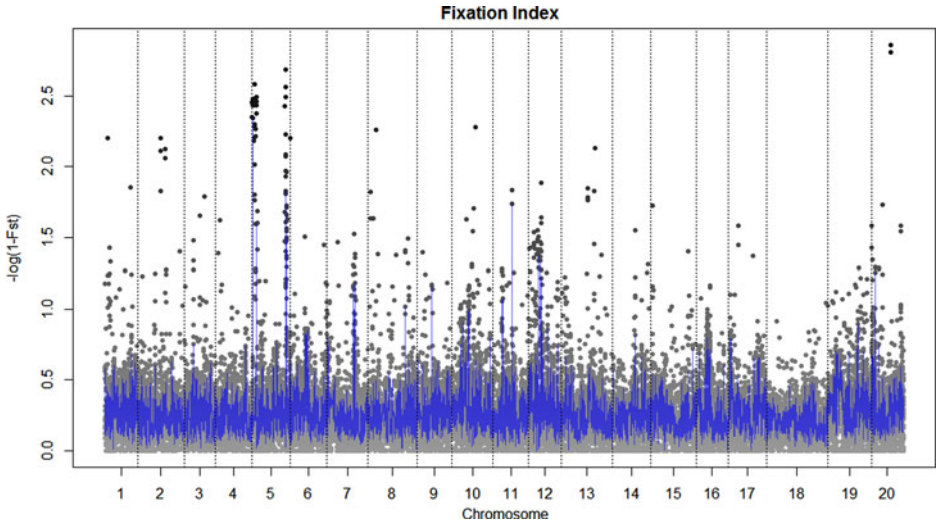


**Fig. 2.** Fixation index of the USDA soybean germplasm collection among eight clusters.

presented by chromosome and with more detail in the Supplementary material (File S1). There are few genomic regions where the smoothed kernel values were above the 99% quantile, and these were likely associated with the causative differentiation of clusters. These regions occurred on chromosomes 5, 7, 8, 9, 11 and 12 (Fig. 2).

Smoothed $F_{ST}$ values higher than 1.45 (99.75% quantile, $F_{ST}$ 0.77) appeared in three regions of chromosome 5, and one region of chromosome 11. The corresponding SNP markers were *Glyma05g10882839* (uncharacterized LOC100 807579), *Glyma05g4381619* (no matching BLAST), *Glyma 05g4852641* (no matching BLAST) and *Glyma11g35 294621* (Glyma11g33480.1, NADH dehydrogenase).

The peaks presented in Fig. 2 are likely associated with alleles that are exclusive to a given cluster or pair of clusters. Table 3 provides a summary of the exclusive alleles found for individual clusters. A total of 914 SNPs were exclusive to specific clusters. Most rare alleles occurred in clusters 3 and 7, whereas cluster 4 did not show any exclusive alleles.

**Table 3.** Number of exclusive alleles and alleles shared between pairs of clusters

| Cluster | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|
| 1 | 7 | 0 | 20 | 2 | 0 | 1 | 24 | 0 |
| 2 | – | 10 | 4 | 0 | 0 | 0 | 20 | 0 |
| 3 | – | – | 372 | 2 | 11 | 9 | 789 | 7 |
| 4 | – | – | – | 0 | 1 | 5 | 12 | 0 |
| 5 | – | – | – | – | 2 | 0 | 72 | 0 |
| 6 | – | – | – | – | – | 5 | 18 | 0 |
| 7 | – | – | – | – | – | – | 516 | 43 |
| 8 | – | – | – | – | – | – | – | 2 |

## *Linkage disequilibrium decay*

Figure 3 presents the LD decay. From the LD, we inferred the overall effective population to be 106.18 individuals, with 95% confidence intervals of {16.84, 195.52}. The effective population size was validated using an empirical approach based on $F_{ST}$ and made available on the Supplementary material (File S1, page 4).

In spite of the large confidence intervals, which are likely associated with genome-duplication (Shoemaker *et al.*, 2006), and the ascertainment bias associated with the SNP array (Lachance and Tishkoff, 2013), this result indicates that an idealized population of approximately 106 unrelated individuals would be enough to comprise all soybean diversity in the collection.

## Discussion

Soybean has undergone many genetic bottlenecks that have narrowed its genetic basis reducing the genetic diversity (Hyten *et al.*, 2006; Tavaud-Pirra *et al.*, 2009; Guo *et al.*, 2010). The study conducted by Hyten *et al.* (2006) involved sequencing of 102 genes from four different populations of soybean before and after the genetic bottlenecks. The study concluded how human-mediated events have weathered the genetic diversity in soybean (Hyten *et al.*, 2006). Similarly, recent advancement in high-throughput sequencing and availability of complete genomic sequence of soybean helped better understand an evolutionary event such as domestication and phenotypic diversification (Hyten *et al.*, 2006; Schmutz *et al.*, 2010; Zhao *et al.*, 2015).

https://doi.org/10.1017/S1479262118000102 Published online by Cambridge University Press

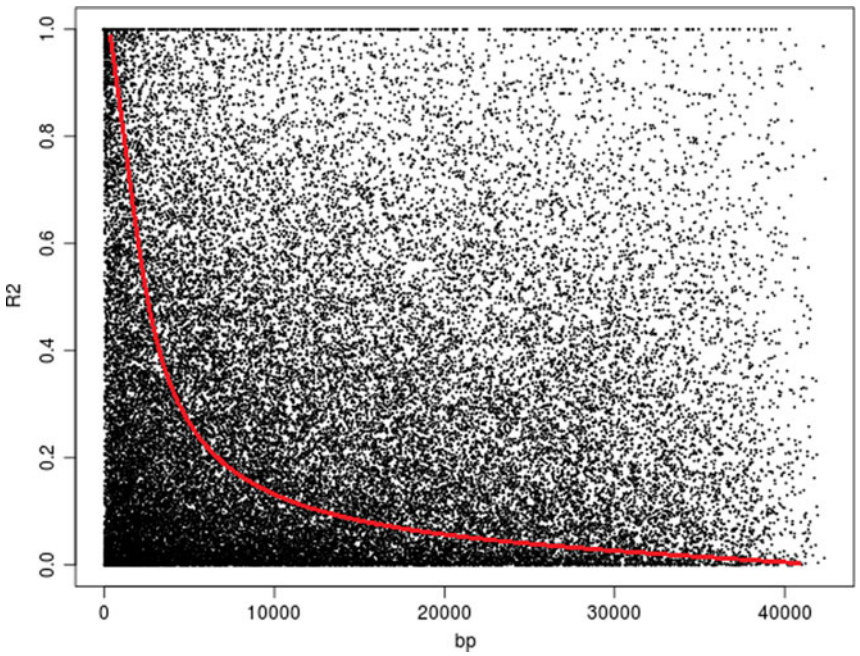


**Fig. 3.** Linkage Disequilibrium decay in the USDA soybean germplasm collection.

Nucleotide fixation driven by selection has reduced the genetic diversity in cultivated soybean when compared with wild progenitors (Zhao *et al.,* 2015), and loss of diversity is mostly associated with domestication rather than genetic improvement (Zhao *et al.,* 2015).

New insights on haplotype diversity in wild and cultivated soybean provided evidence of evolutionary events affecting the population dynamics and helping soybean researchers in their pursuit of crop improvement (Zhao *et al.,* 2015). Due to limited diversity in soybean, breeders and researchers have been introgressing rare alleles from undomesticated material to expand the genetic base, thus preferred over other sources of genetic variabilities, such as mutation breeding (Ha *et al.,* 2014) and transformation (Yamada *et al.,* 2012). Introgression of exotic alleles was not limited to *G. soja*, also encompassing perennial *Glycine* species (Sherman-Broyles *et al.,* 2014), such as *G. tomentella* that has been recently exploited and found to be a valuable resource (Singh and Nelson, 2015).

## Genetic variability

An effective population size of approximately 106 accessions reinforces the concept of a narrow genetic base for soybean reported in previous studies performed on a smaller scale (Zhu *et al.,* 2003; Li *et al.,* 2013; Min *et al.,* 2013). From the breeding perspective, this restricted genetic variability imposes constraints on long-term genetic gains (Henryon *et al.,* 2014).

Song *et al.* (2015) reported that the USDA soybean germplasm collection has redundant accessions, with approximately 30% of domesticated and 23% of undomesticad lines being nearly identical to other entries. Such similarity can be attributed to highly related lines (i.e. as near-isogenic lines), mutation-driven populations that differ in point-specific SNPs, or mislabelling resulting in duplicate accessions.

Three major bottlenecks may have restricted genetic diversity outside the centre of origin (Hyten *et al.,* 2006), comprising initial domestication, the introduction of few landraces into novel territories and local selective breeding. The high number of alleles exclusive between clusters 3 and 7 (Table 3) suggests that tropical and temperate domestication of soybeans may have occurred separately. Yet, much of the existing genetic diversity of soybeans still resides in germplasm that preceded the bottlenecks (He *et al.,* 2012; Zhao *et al.,* 2015), including landraces and undomesticated material.

Surprisingly, the genetic correlation between grain yield and domestication is weak (Table 2), indicating that exotic alleles from wild relatives mostly do not affect seed production. QTL have been successfully introgressed from G. *soja* to *G. max* (Concibido *et al.*, 2003; Wang, *et al.,* 2004; Wang *et al.,* 2010), although backcrosses are needed for a fully agronomically desirable phenotype (Ertl and Fehr, 1985).

## Differentiation

Many genomic regions are associated with the underlying genetic basis of differentiation between the eight major

https://doi.org/10.1017/S1479262118000102 Published online by Cambridge University Press

clusters (Fig. 2, Suppl. File S1 page 5), and many alleles occur exclusively in specific clusters or pairs of clusters (Table 3).

The domestication of soybean occurred distinctly multiple times, and such events have been detected in both nucleic and cytoplasmic DNA (Xu *et al.,* 2002; Schmutz *et al.,* 2010). Some function regions underwent fixation because soybeans were selected for specific purposes (Hou *et al.,* 2009; Shi *et al.,* 2010), while non-functional paralogue regions originated from genome duplications (Schmutz *et al.,* 2010) were often fixed due to random genetic drift at separate points in time (Recker *et al.,* 2013; Zhao *et al.,* 2015). This distinction of selection and drift is particularly informative with regard to the time of differentiation and genetic structure.

Association studies based fixation index aiming to reveal the genetic nature of soybean domestication (Narvel *et al.,* 2000) can identify which genes, their function and location, characterize wild relatives, landrace strains and elite material (Hyten *et al.,* 2006; Wang *et al.,* 2014; Song *et al.,* 2015). Alleles associated with domestication may have been lost in bottlenecks at random or may be linked to agronomic traits of importance (Zhao *et al.,* 2015). The development of experimental populations would be necessary to truly distinguish drift from domestication QTL (Recker *et al.,* 2013).

An interesting association between grain yield and seed weight (0.46) is presented in Table 2, where seed weight also appears to be more heritable than grain yield (0.62 > 0.52). This result indicates that seed weight is a good candidate trait to be taken into account when incorporating novel sources of germplasm into breeding programmes. However, this trend has not been observed in experimental populations from elite germplasm (Xavier *et al.,* 2017).

## *Quantitative control of traits*

As expected, the estimates of heritability at the whole-germplasm level (Table 2) were high, reflecting a predominantly additive genetic control of the quantitative traits under evaluation. This phenomenon occurs due to the large genetic variance in comparison with the noise from phenotypic data, boosting the signal-to-noise ratio (Muir, 2007). In addition, the magnitude of the genomic associations among traits, here represented by genetic correlations, also depends on the genetic signalling of individual traits, as represented by the heritability.

Estimates of covariance components affect the downstream estimation of breeding values (Searle, 1961) and, subsequently, the efficacy of multi-trait selection indices (Hazel, 1943). The parameters presented in Table 2 provide insight into the general trend of correlated response to selection in soybeans. Although the abundance of genetic variation could inflate the coefficients, our estimates of heritability and genetic correlation from the germplasm collection agree with previous reports from elite panels, random mating populations and experimental populations (Johnson *et al.,* 1955; Kwon and Torrie, 1964; Ecochard and Ravelomanantsoa, 1982; Arshad *et al.,* 2006; Recker *et al.,* 2014).

## *Consideration of maturity*

Some traits show fairly stable genetic control. For instance, the strong negative genetic correlation between protein and oil (−0.83) was similar to previous studies, which reported −0.75 (Recker *et al.,* 2014) and −0.66 (Kwon and Torrie, 1964). The association between traits is can also be population specific. The association between flowering date and maturity date (0.85) in the germplasm was much stronger than in the random mating population (0.44) presented by Recker *et al.* (2014) or the elite panel (0.34) presented by Arshad *et al.* (2006), but was consistent with the wide-maturity bi-parental population (0.87) presented by Ecochard and Ravelomanantsoa (1982). This discrepancy in the genetic correlation between flowering and maturity among the various populations is directly related to the genetic variability of each panel. Populations ranging across multiple maturity groups are more likely to display larger genetic correlations for traits that control the reproductive cycle.

The maturity groups of soybean are defined by few major genes (Molnar *et al.,* 2003; Samanfar *et al.,* 2017), which affects the heritability estimate (0.85). Nevertheless, we did not observe the same level of additive genetic control for maturity date (0.38). Besides the underlying genetic control of maturity that defines 'what grows where', the trait maturity date also relies on the planting date and environmental factors defined by the location in which the accessions were evaluated, factors such as temperature and photoperiod (Tasma *et al.,* 2001; Xu *et al.,* 2013). These factors make the flowering and maturity date traits not directly comparable across a wide range of maturity groups in situations when not all the germplasm collection is evaluated at the same growing site.

## Conclusions

Using the passport data and genomic information available, we were able to summarize population parameters in the USDA soybean germplasm collection, finding the existence of eight clear subpopulations. $F_{ST}$ tracked regions that could explain the differentiation among clusters. The high levels of LD found to indicate that the effective population size of the germplasm collection is approximately 106 individuals. We also presented and discussed the quantitative merit of major traits.

This study provides a summary of various population and quantitative parameters. Such information is particularly

https://doi.org/10.1017/S1479262118000102 Published online by Cambridge University Press

valuable for the pre-breeding processes that seek to expand the genetic basis of breeding germplasm. Here we also provide phenotypes and genotypes (and clusters) in the supplementary material to facilitate breeders using the germplasm information. Maturity and seed weight appeared as two key traits when selecting material from the germplasm collection.

Future research could focus on the genomic regions or the candidate genes that have been identified as associated with important traits, such as yield, plant height, seed weight and lodging. Further studies could be performed *in silico* using the SoyNAM public dataset (Xavier *et al.*, 2018) to contrast or complement genetic investigations on the USDA germplasm collection. We believe that new studies should also seek robust phenotypic characterization of the valuable quantitative traits due to the tremendous amount of information that phenotypic diversity could provide for researchers to improve soybean.

## Supplementary material

The supplementary material for this article can be found at https://doi.org/10.1017/S1479262118000102

## Acknowledgements

We are grateful to Shogo Tsuruta for providing the software GIBBS3F90, which was able to support enough entries. We also thank Randall Nelson for providing the passport data.

## Declarations

## Authors' contribution

AX wrote the manuscript with contributions from RT. AX conducted the statistical analyses. WMM and KMR provided the general objectives and an important insight into the analyses. All authors read and reviewed the manuscript.

## References

Akey JM, Zhang G, Zhang K, Jin L and Shriver MD (2002) Interrogating a high-density SNP map for signatures of natural selection. *Genome Research* 12: 1805–1814.

Arshad MU, Ali N and Ghafoor A (2006) Character correlation and path coefficient in soybean *Glycine max* (L.) Merrill. *Pakistan Journal of Botany* 38: 121.

Bandillo N, Jarquin D, Song Q, Nelson R, Cregan P, Specht J and Lorenz A (2015) A population structure and genome-wide association analysis on the USDA soybean germplasm collection. *Plant Gene* 8: 1–13. doi: 10.3835/plantgenome2015.04.0024.

Brown-Guedira GL, Thomson JA, Nelson RL and Warburton ML (2000) Evaluation of genetic diversity of soybean introductions and North American ancestors using RAPD and SSR markers. *Crop Science* 40: 815–823.

Carter TE, Hymowitz T and Nelson RL (2004a) Biogeography, local adaptation, Vavilov and genetic diversity in soybean. In: Werner D (eds) *Biological Resources and Migration*. Berlin: Springer, pp. 47–59.

Carter TE, Nelson R, Sneller CH and Cui Z (2004b) In soybeans: improvement, production, and uses. In: Boerma HR and Specht JE (eds) *Vol Agronomy*. Madison, WI: American Society of Agronomy, Crop Science Society of America, Soil Science Society of America, no 16, pp. 303–416.

Chan C, Qi X, Li M-W, Wong F-L and Lam H-M (2012) Recent developments of genomic research in soybean. *Journal of Genetics* and *Genomics* 39: 317–324.

Chang H, Lipka AE, Domier LL and Hartman GL (2016) Characterization of disease resistance loci in the USDA soybean germplasm collection using genome-wide association studies. *Phytopathology* 106: 1139–1151.

Concibido V, La Vallee B, Mclaird P, Pineda N, Meyer J, Hummel L, Yang J, Wu K and Delannay X (2003) Introgression of a quantitative trait locus for yield from *Glycine soja* into commercial soybean cultivars. *Theoretical and Applied Genetics* 106: 575–582.

Cox TF and Cox MA (2000) *Multidimensional Scaling*. CRC Press.

DeJong G and VanNoordwijk: AJ (1992) Acquisition and allocation of resources: genetic (co) variances, selection, and life histories. *American Naturalist* 139: 749–770.

Doebley JF, Gaut BS and Smith BD (2006) The molecular genetics of crop domestication. *Cell* 127: 1309–1142.

Ecochard R and Ravelomanantsoa Y (1982) Genetic correlations derived from full-sib relationships in soybean (*Glycine max* Merr.). *Theoretical and Applied Genetics* 63: 9–15.

Ertl DS and Fehr WR (1985) Agronomic performance of soybean genotypes from *Glycine max* x *Glycine soja* crosses. *Crop Science* 25: 589–592.

Flori L, Fritz S, Jaffrézic F, Boussaha M, Gut I, Heath S, Foulley JL and Gautier M (2009) The genome response to artificial selection: a case study in dairy cattle. *PLoS ONE* 4: e6595.

Grant D, Nelson RT, Cannon SB and Shoemaker RC (2009) Soybase, the USDA-ARS soybean genetics and genomics database. *Nucleic Acids Research* 38: D843–D846.

Guo J, Wang Y, Song C, Zhou J, Qiu L, Huang H and Wang Y (2010) A single origin and moderate bottleneck during domestication of soybean (*Glycine max*): implications from microsatellites and nucleotide sequences. *Annals of Botany* 106: 505–514.

Ha BK, Lee KJ, Velusamy V, Kim JB, Kim SH, Ahn JW, Kang SY and Kim DS (2014) Improvement of soybean through radiation-induced mutation breeding techniques in Korea. *Plant Genetic Resources* 12: S54–S57.

Hazel LN (1943) The genetic basis for constructing selection indexes. *Genetics* 28: 476–490.

He S, Wang Y, Volis S, Li D and Yi T (2012) Genetic diversity and population structure: implications for conservation of wild soybean (*Glycine soja* Sieb. et Zucc) based on nuclear and chloroplast microsatellite variation. *International Journal of Molecular Sciences* 13: 12608–12628.

Henryon M, Berg P and Sørensen AC (2014) Animal-breeding schemes using genomic information need breeding plans designed to maximise long-term genetic gains. *Livestock Science* 166: 38–47.

Holsinger KE and Weir BS (2009) Genetics in geographically structured populations: defining, estimating and interpreting FST. *Nature Reviews Genetics* 10: 639–650.

Hou A, Chen P, Alloatti J, Li D, Mozzoni L, Zhang B and Shi A (2009) Genetic variability of seed sugar content in worldwide soybean germplasm collections. *Crop Science* 49: 903–912.

Hymowitz T (2008) The history of the soybean. In Johnson L, White PJ and Galloway R (eds) *Soybeans: Chemistry, Production, Processing and Utilization*. Urbana, IL: AOCS Press, pp. 1–32.

Hyten DL, Song Q, Zhu Y, Choi I, Nelson RL, Costa JM, Specht JE, Shoemaker RC and Cregan PB (2006) Impacts of genetic bottlenecks on soybean genome diversity. *Proceedings of the National Academy of Sciences of the United States opf America* 103: 16666–16671.

James G, Witten D, Hastie T and Tibshirani R (2013) *An Introduction to Statistical Learning*. New York: Springer, 1st ed. 2013, Corr. 5th printing 2015 Edition.

Jarquin D, Specht J and Lorenz A (2016) Prospects of genomic prediction in the USDA soybean germplasm collection: historical data creates robust models for enhancing selection of accessions. *G3: Genes| Genomes| Genetics* 6: 2329–2341.

Johnson HW, Robinson HF and Comstock RE (1955) Estimates of genetic and environmental variability in soybeans. *Agronomy Journal* 47: 314–318.

Jombart T and Ahmed I (2011) Adegenet 1.3–1: new tools for the analysis of genome-wide SNP data. *Bioinformatics* 27: 3070–3071.

Kuroda Y, Kaga A, Tomooka N, Yano H, Takada Y, Kato S and Vaughan D (2013) QTL affecting fitness of hybrids between wild and cultivated soybeans in experimental fields. *Ecology and Evolution* 3: 2150–2168. http://doi.org/10.1002/ece3.606.

Kwon SH and Torrie JH (1964) Heritability and interrelationship among traits of two soybean populations. *Crop Science* 4: 196–198.

Lachance J and Tishkoff SA (2013) SNP ascertainment bias in population genetic analyses: why it is important, and how to correct it. *Bioessays* 35: 780–786.

Li YH, Li W, Zhang C, Yang L, Chang RZ, Gaut BS and Qiu LJ (2010) Genetic diversity in domesticated soybean (*Glycine max*) and its wild progenitor (*Glycine soja*) for simple sequence repeat and single-nucleotide polymorphism loci. *New Phytologist* 188: 242–253.

Li YH, Zhao SC, Ma JX, Li D, Yan L, Li J, Qi XT, Guo XS, Zhang L, He WM and Chang RZ (2013) Molecular footprints of domestication and improvement in soybean revealed by whole genome re-sequencing. *BMC Genomics* 14: 579.

Mardia KV (1978) Some properties of classical multidimensional scaling. *Communications on Statistics – Theory and Methods* A7: 1233–1241.

Min W, Run-zhi L, Wan-ming Y and Wei-jun D (2013) Assessing the genetic diversity of cultivars and wild soybeans using SSR markers. *African Journal of Biotechnology* 9: 4857–4866.

Misztal I, Tsuruta S, Strabel T, Auvray B, Druet T and Lee DH (2002) BLUPF90 and related programs (BGF90). In Proceedings of the 7th World Congress on Genetics Applied to Livestock Production, Montpellier, France, August 2002; Session 28. (pp. 1–2). Institut National de la Recherche Agronomique (INRA).

Molnar SJ, Rai S, Charette M and Cober ER (2003) Simple sequence repeat (SSR) markers linked to E1, E3, E4, and E7 maturity genes in soybean. *Genome* 46: 1024–1036.

Muir WM (2007) Comparison of genomic and traditional BLUP-estimated breeding value accuracy and selection response under alternative trait and genomic parameters. *Journal of Animal Breeding and Genetics* 124: 342–355.

Murtagh F and Legendre P (2014) Ward's hierarchical agglomerative clustering method: which algorithms implement ward's criterion? *Journal of Classification* 31: 274–295.

Narvel JM, Fehr WR, Chu WC, Grant D and Shoemaker RC (2000) Simple sequence repeat diversity among soybean plant introductions and elite genotypes. *Crop Science* 40: 1452–1458.

Nei M (1972) Genetic distance between populations. *American Naturalist* 106: 283–292.

Paradis E, Claude J and Strimmer K (2004) APE: analyses of phylogenetics and evolution in R language. *Bioinformatics* 20: 289–290.

Recker JR, Burton JW, Cardinal A and Miranda L (2013) Analysis of quantitative traits in two long-term randomly mated soybean populations: I. *Genetic Variances* 53: 1375–1383.

Recker JR, Burton JW, Cardinal A and Miranda L (2014) Genetic and phenotypic correlations of quantitative traits in two long-term, randomly mated soybean populations. *Crop Science* 54: 939–943.

Reif JC, Melchinger AE and Frisch M (2005) Genetical and mathematical properties of similarity and dissimilarity coefficients applied in plant breeding and seed bank management. *Crop Science* 45: 1–7.

Samanfar B, Molnar SJ, Charette M, Schoenrock A, Dehne F, Golshani A, Belzile F and Cober ER (2017) Mapping and identification of a potential candidate gene for a novel maturity locus, E10, in soybean. *Theoretical and Applied Genetics* 130: 377–390.

Schmutz J, Cannon SB, Schlueter J, Ma J, Mitros T, Nelson W, Hyten DL, Song Q, Thelen JJ, Cheng J and Xu D, (2010) Genome sequence of the palaeopolyploid soybean. *Nature* 463: 178–183.

Searle SR (1961) Phenotypic, genetic and environmental correlations. *Biometrics* 17: 474–480.

Sherman-Broyles S, Bombarely A, Powell AF, Doyle JL, Egan AN, Coate JE and Doyle JJ (2014) The wild side of a major crop: soybean's perennial cousins from down under. *American Journal of Botany* 101: 1651–1665.

Shi A, Chen P, Zhang B and Hou A (2010) Genetic diversity and association analysis of protein and oil content in food-grade soybeans from Asia and the United States. *Plant Breeding* 129: 250–256.

Shoemaker RC, Schlueter J and Doyle JJ (2006) Paleopolyploidy and gene duplication in soybean and other legumes. *Current Opinion in Plant Biology* 9: 104–109.

Singh RJ and Hymowitz T (1989) The genomic relationships among *Glycine soja* Sieb. and Zucc. *G. max* (L.) Merr. and '*G. gracilis*' Skvortz. *Plant Breeding* 103: 171–173.

Singh RJ and Nelson RL (2015) Intersubgeneric hybridization between *Glycine max* and *G. tomentella*: production of F1, amphidiploid, BC1, BC2, BC3, and fertile soybean plants. *Theoretical and Applied Genetics* 128: 1117–1136.

Slatkin M and Excoffier L (1996) Maximization algorithm. *Heredity* 76: 377–383.

Song Q, Hyten DL, Jia G, Quigley CV, Fickus EW, Nelson RL and Cregan PB (2013) PB. Development and evaluation of SoySNP50K, a high-density genotyping array for soybean. *PLoS ONE* 8: e54985.

Song Q, Hyten DL, Jia G, Quigley CV, Fickus EW, Nelson RL and Cregan PB (2015) Fingerprinting soybean germplasm and its utility in genomic research. *G3: Genes| Genomes| Genetics* 5: 1999–2006.

Stekhoven DJ and Buhlmann P (2012) Missforest: non-parametric missing value imputation for mixed-type data. *Bioinformatics* 28: 112–118.

Stranden I and Christensen OF (2011) Allele coding in genomic evaluation. *Genetics Selection Evolution* 43: 1–11.

Sved JA, Cameron EC and Gilchrist AS (2013) Estimating effective population size from linkage disequilibrium between unlinked loci: theory and application to fruit fly outbreak populations. *PLoS ONE* 8: e69078.

Tasma IM, Lorenzen LL, Green DE and Shoemaker RC (2001) Mapping genetic loci for flowering time, maturity, and photoperiod insensitivity in soybean. *Molecular Breeding* 8: 25–35.

Tavaud-Pirra M, Sartre P, Nelson R, Santon S, Texier N and Roumet P (2009) Genetic diversity in a soybean collection. *Crop Science* 49: 895–902.

Wang D, Graef GL, Procopiuk AM and Diers BW (2004) Identification of putative QTL that underlie yield in interspecific soybean backcross populations. *Theoretical and Applied Genetics* 108: 458–467.

Wang KJ, Li XH, Zhang JJ, Chen H, Zhang ZL and Yu GD (2010) Natural introgression from cultivated soybean (*Glycine max*) into wild soybean (*Glycine soja*) with the implications for origin of populations of semi-wild type and for biosafety of wild species in China. *Genetic Resources and Crop Evolution* 57: 747–761.

Wang Y, Lu J, Chen S, Shu L, Palmer RG, Xing G, Li Y, Yang S, Yu D, Zhao T and Gai J, (2014) Exploration of presence/absence variation and corresponding polymorphic markers in soybean genome. *Journal of Integrative Plant Biology* 56: 1009–1019.

Waples RS, Antao T and Luikart G (2014) Effects of overlapping generations on linkage disequilibrium estimates of effective population size. *Genetics* 197: 769–780.

Weir BS and Cockerham CC (1984) Estimating F-statistics for the analysis of population structure. *Evolution* 38: 1358–1370.

Wen Z, Ding Y, Zhao T and Gai J (2009) Genetic diversity and peculiarity of annual wild soybean (*G. soja* Sieb. et Zucc.) from various eco-regions in China. *Theoretical and Applied Genetics* 119: 371–381.

Wright S (1949) The genetical structure of populations. *Annals of Eugenics* 15: 323–354.

Wright S (1965) The interpretation of population structure by F-statistics with special regard to systems of mating. *Evolution* 19: 395–420.

Xavier A, Xu S, Muir WM and Rainey KM (2015) NAM: association studies in multiple populations. *Bioinformatics* 31: 3862–3864.

Xavier A, Muir WM, Craig B and Rainey KM (2016) Walking through the statistical black boxes of plant breeding. *Theoretical and Applied Genetics* 129: 1933–1949.

Xavier A, Hall B, Casteel S, Muir W and Rainey KM (2017) Using unsupervised learning techniques to assess interactions among complex traits in soybeans. *Euphytica* 213: 200.

Xavier A, Jarquin D, Howard R, Ramasubramanian V, Specht JE, Graef GL, Beavis WD, Diers BW, Song Q, Cregan PB and Nelson R (2018) Genome-Wide analysis of grain yield stability and environmental interactions in a multiparental soybean population. *G3: Genes, Genomes, Genetics* 8: 519–529.

Xu D, Abe J, Gai J and Shimamoto Y (2002) Diversity of chloroplast DNA SSRs in wild and cultivated soybeans: evidence for multiple origins of cultivated soybean. *Theoretical and Applied Genetics* 105: 645–653.

Xu M, Xu Z, Liu B, Kong F, Tsubokura Y, Watanabe S, Xia Z, Harada K, Kanazawa A, Yamada T and Abe J (2013) Genetic variation in four maturity genes affects photoperiod insensitivity and PHYA-regulated post-flowering responses of soybean. *BMC Plant Biology* 13: 1.

Yamada T, Takagi K and Ishimoto M (2012) Recent advances in soybean transformation and their application to molecular breeding and genomic analysis. *Breeding Science* 61: 480–494.

Yamamichi M and Innan H (2012) Estimating the migration rate from genetic variation data. *Heredity* 108: 362.

Zera AJ and Harshman: LG (2001) The physiology of life history trade-offs in animals. *Annual Review of Ecology and Systematics* 32: 95–126.

Zhang J, Song Q, Cregan PB and Jiang GL (2016) Genome-wide association study, genomic prediction and marker-assisted selection for seed weight in soybean (*Glycine max*). *Theoretical and Applied Genetics* 129: 117–130.

Zhao S, Zheng F, He W, Wu H, Pan S and Lam HM (2015) Impacts of nucleotide fixation during soybean domestication and improvement. *BMC Plant Biology* 15: 81.

Zhou Z, Jiang Y, Wang Z, Gou Z, Lyu J, Li W, Yu Y, Shu L, Zhao Y, Ma Y and Fang C (2015) Resequencing 302 wild and cultivated accessions identifies genes related to domestication and improvement in soybean. *Nature Biotechnology* 33: 408–414.

Zhu YL, Song QJ, Hyten DL, Van Tassell CP, Matukumalli LK, Grimm DR, *et al*. (2003) Single-nucleotide polymorphisms in soybean. *Genetics* 163: 1123–1134.