Heterosis, inbreeding depression and genetic divergence in maize

Fernando Silva Aguilar

Iowa State University

Follow this and additional works at: https://lib.dr.iastate.edu/etd

Recommended Citation

Silva Aguilar, Fernando, "Heterosis, inbreeding depression and genetic divergence in maize" (2020). Graduate Theses and Dissertations. 18229.
https://lib.dr.iastate.edu/etd/18229

This Dissertation is brought to you for free and open access by the Iowa State University Capstones, Theses and Dissertations at Iowa State University Digital Repository. It has been accepted for inclusion in Graduate Theses and Dissertations by an authorized administrator of Iowa State University Digital Repository. For more information, please contact digirep@iastate.edu.
Heterosis, inbreeding depression and genetic divergence in maize

by

Fernando Silva Aguilar

A dissertation submitted to the graduate faculty
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Major: Plant Breeding

Program of Study Committee:
Jode Edwards, Co-major Professor
Kendall Lamkey, Co-major Professor
Paul Scott
Jianming Yu
Dan Nettleton

The student author, whose presentation of the scholarship herein was approved by the program of study committee, is solely responsible for the content of this dissertation. The Graduate College will ensure this dissertation is globally accessible and will not permit alterations after a degree is conferred.

Iowa State University
Ames, Iowa
2020

Copyright © Fernando Silva Aguilar, 2020. All rights reserved.
DEDICATION

I dedicate this dissertation to my beloved mother María Gladis Aguilar Moreno, my brother Bernardo Silva Aguilar, and to my wife Adriana Vega Osorio.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>LIST OF FIGURES</th>
<th>v</th>
</tr>
</thead>
<tbody>
<tr>
<td>LIST OF TABLES</td>
<td>vii</td>
</tr>
<tr>
<td>ACKNOWLEDGMENTS</td>
<td>viii</td>
</tr>
<tr>
<td>ABSTRACT</td>
<td>ix</td>
</tr>
</tbody>
</table>

## CHAPTER 1. GENERAL INTRODUCTION .......................................................... 1
- Hybrid-seed production system .................................................. 1
- Heterosis ................................................................................. 3
- Inbreeding and inbreeding depression ........................................... 5
- Quantitative genetic studies of heterosis and inbreeding .................. 7
- Detection of identity-by-descent (IBD) segments with genome-wide genotypic data .......................................................... 8
- Objectives .............................................................................. 11
- Dissertation organization .......................................................... 12

## CHAPTER 2. POPULATION DIVERGENCE AND INBREEDING DEPRESSION IN MAIZE HETEROSIS ............................................................ 13
- Abstract .................................................................................... 13
- Introduction ............................................................................. 14
- Materials and methods ............................................................. 18
  - Genetic materials ................................................................. 18
  - Experimental design .............................................................. 18
  - Data collection ...................................................................... 19
  - Data analysis .......................................................................... 20
  - Genetic model ......................................................................... 21
  - Heterosis ............................................................................... 25
  - Average inbreeding depression in the parents (ID) ....................... 26
- Results and discussion ............................................................... 27
  - Heterosis and average inbreeding depression in parents ............... 27
  - Genetic importance of a hybrid seed production system ................ 29
- Summary .................................................................................... 30
- Tables and Figures ..................................................................... 31
- References ................................................................................. 38

## CHAPTER 3. GENETIC COMPOSITION AND GENETIC DIVERSITY OF 44 EXPIRED PLANT VARIETY PROTECTION (ex-PVP) MAIZE INBRED LINES THROUGH HIGH-RESOLUTION DETECTION OF IDENTITY BY DESCENT SEGMENTS ........................................... 42
- Abstract .................................................................................... 42
- Introduction ............................................................................. 43
- Materials and methods ............................................................. 48
  - Plant material ....................................................................... 48
Figure 2.1. Mating scheme for the panmictic populations (P1 and P2) with average inbreeding compared to a metapopulation bigger than zero (FST > 0). Panmictic populations are inbred for several generations to create a population of inbreds. Genetically, crossing the two populations of inbreds will produce the same F1 as crossing the two panmictic populations. Random mating individuals in the F1 gives origin to the F2, considered as the reference population. ............................ 36

Figure 2.2. Simple linear regression of midparent heterosis versus midparent values for all crosses and A) grain yield, B) plant height, and C) ear height. Dots correspond to inbred-by-inbred crosses, upward triangle to population-by-inbred crosses, and the downward triangles to population-by-population crosses. .......................... 37

Figure 2.3. Simple linear regression of inbreeding depression in the parent populations versus midparent heterosis for A) grain yield, B) plant height, and C) ear height. Upward triangles represent the population-by-inbred cross, while downward triangles represent the population-by-population crosses. ............................. 38

Figure 3.1. Genetic diversity within and between the stiff stalk (SSS) and non-stiff stalk (NSS) heterotic groups. Diversity was estimated as 1 minus the kinship coefficient calculated with A) a marker-based and B) a pedigree-based distance matrix. Horizontal black line in the box and red diamond corresponds to the median and mean diversity per group, respectively. ................................................................. 72

Figure 3.2. Dendrogram based on UPGMA clustering from the marker-based genetic distance matrix computed from the posterior probability of IBD status at each locus. This circular dendrogram highlights the heterotic subgroup structure found for the 52 inbred lines. Major heterotic subgroups are labeled broadly based on their relevant ancestor line. ........................................................................................................... 73

Figure 3.3. Dendrogram based on UPGMA clustering from a kinship matrix computed with the VanRaden’s (2008) methodology. This circular dendrogram highlights the heterotic subgroup structure found for the 52 inbred lines. Major heterotic subgroups are labeled broadly based on their relevant ancestor line. ......................... 74

Figure 3.4. Dendrogram based on UPGMA clustering from the pedigree-based genetic distance matrix. This circular dendrogram highlights the heterotic subgroup structure found for the 52 inbred lines. Major heterotic subgroups are labeled broadly based on their relevant ancestor line. ................................................................. 75
Figure 3.5. Proportion of the genome classified as IBD between ex-PVP lines and each of the key ancestor lines (B14, B37, B73, PHJ40, PHG39, PHK29, LH132, Mo17, Oh43, and PH207) for each heterotic subgroup identified with marker-based dissimilarity values. Non-IBD category (grey bars) correspond to SNPs not included within an IBD segment between the line and key ancestor. .......................... 76

Figure 3.6. Histogram showing the density distribution in logarithmic scale (log10[Mb]) of the identity by descent segments detected among the key ancestor lines and ex-PVP lines within A) stiff stalk (SSS) and B) non-stiff stalk (NSS) heterotic group. The red dashed lines represent the quantiles 25% (Q1) and 75% (Q3). The grey area represents the density probability for IBD segment lengths. Inside box is the summary statistic with Q2 representing the median. .......................... 77

Figure 3.7. Distribution of the number of generations of recombination an IBD segment was originated for the stiff stalk (SSS) and non-stiff stalk (NSS) heterotic group. The curly bracket corresponds to the cumulative percentage for inbreeding created between 35 and 135 generations. ............................................................. 77

Figure 4.1. Values for A) the statistic Delta K for a model-based clustering algorithm and B) the Bayesian Information Criterion (BIC) for a principal component analysis (PCA) to identify the best K within the 487 doubled haploid lines. K corresponds to the number of subpopulations. Delta K statistic detects the rate of change between successive K values, while BIC detects the fit of the model when K varies from 1 to 10. The highest delta K value corresponds to the true value for K, while the lowest BIC corresponds to the best model according to the data provided. ................................................................................. 121

Figure 4.2. Quantile-quantile plots for plant height based on observed versus expected −log10(P values) for the CMLM, FarmCPU, FaST-LMM, MLMM, and SUPER methods when the A) IBDLD and B) G kinship matrix were used. ........ 122

Figure 4.3. Comparison of significant SNPs associated with plant height between SUPER and FarmCPU using both the VanRaden (2008) (G Matrix) and IBDLD kinship matrices................................................................. 123
# LIST OF TABLES

<table>
<thead>
<tr>
<th>Table 2.1. Derivation process and development of the fourteen (14) parents evaluated in the experiment.</th>
<th>31</th>
</tr>
</thead>
<tbody>
<tr>
<td>Table 2.2. Crosses made between the synthetic populations, between synthetic populations and the inbred line B129, and between inbred lines.</td>
<td>32</td>
</tr>
<tr>
<td>Table 2.3. Genetic expected means for each generation, midparent value at panmixia (MPV) and complete inbreeding (IMPV), and heterosis for panmictic populations and for inbred-by-inbred crosses.</td>
<td>32</td>
</tr>
<tr>
<td>Table 2.4. Bayesian information criteria (BIC) for the 24 possible linear models. The models provided contained all fixed effects described in the materials and methods section plus the random effects containing an ‘X’ in the corresponding column. The lowest BIC value is denoted with an asterisk.</td>
<td>33</td>
</tr>
<tr>
<td>Table 2.5. Mean values (standard error) for midparent value (MPV), inbred midparent value (IMPV), midparent heterosis (MH), inbred-midparent heterosis (IMH), midparent F₂ (MF₂H) heterosis, and inbred-midparent F₂ (IMF₂H) heterosis for ear height (cm), plant height (cm), and grain yield (Mg ha⁻¹) for the inbred-by-inbred (IxI), population-by-inbred (PxI), and population-by-population (PxP) crosses.</td>
<td>34</td>
</tr>
<tr>
<td>Table 3.1. List of the 44 ex-PVP and 8 key ancestors lines evaluated, with genetic background, patent applicant company or institution, patent number, number of generations the pedigree is available (Generations), and dataset from where genotypic data was collected.</td>
<td>69</td>
</tr>
<tr>
<td>Table 3.2. Degree of relationship between each key ancestor and ex-PVP per heterotic group. The kinship was calculated with a marker-based (an HMM framework for IBD), G-based (IBS kinship computed with VanRaden (2008) algorithm), and a pedigree-based kinship method.</td>
<td>71</td>
</tr>
<tr>
<td>Table 4.1. List of all significant SNP markers associated with plant height found with FarmCPU, and SUPER when using a kinship matrix computed from IBD segments (IBDLD) or IBS markers (G Matrix).</td>
<td>120</td>
</tr>
</tbody>
</table>
ACKNOWLEDGMENTS

I would like to thank my advisor, Jode Edwards Warren, for his patience and guidance during my Ph.D. study, and to the rest of my committee members for their guidance. Additionally, I am grateful to my family, friends, faculty, and staff for making my time at Iowa State University a wonderful experience. Finally, I thank to COLCIENCIAS Colombia, Fulbright, and LASPAU for all the economic support.
Heterosis and inbreeding depression are known to be opposite phenomena that depend on allele frequencies and directional dominance. Heterosis refers to the superiority of the hybrid over its parents by an increase in the mean of crossbred individuals, while inbreeding depression refers to the reduction in the phenotypic mean of a population. Heterosis has been described as a function of the squared of the genetic divergence in allele frequency of the parents ($\Delta^2$), and dominance ($d$), while inbreeding depression depends on $d$ and allele frequencies. We derived a model of heterosis based on genetic divergence in allele frequencies between the parents ($\Delta$), dominance, and inbreeding depression. Similarly, to better understand $\Delta$ and inbreeding depression we estimate shared identical by descent (IBD) segments between inbred lines of maize. The main objective was to understand the underlying basis of heterosis and to estimate genetic diversity and progenitor’s genetic contribution based on the amount of shared IBD segments. To describe heterosis, six synthetic maize populations and eight inbred lines were used. Three crosses between synthetic populations, three between synthetic populations and the B129 inbred line, and six between inbred lines were evaluated in nine environments under a modified split-plot design with three replications. For easy deductions, heterosis model was defined under a single-locus two-alleles case and tested using a “goodness-of-fit” test. To estimate IBD segments, a set of 44 ex-PVP lines along with eight key ancestors of maize in the U.S. Corn Belt were selected. Shared IBD segments were identified by using a probabilistic approach based on a Hidden Markov Model (HMM) framework. Genetic diversity between individuals was estimated as 1 minus the kinship coefficient. Genome-wide kinship coefficients were calculated from the posterior probability of the IBD status at each locus. Our results showed that a single-locus two allele model of heterosis was adequate to describe the variation in
the mean of each generation and to predict heterosis. Heterosis estimates were significantly higher for crosses involving inbred parent populations, with a limiting value when the parents reach complete inbreeding. Both model and empirical evidence of heterosis shows that population divergence in allele frequency between parents ($\Delta$) is the key driver of heterosis, but that this divergence is achieved when the level of heterozygosis within each population decreases (i.e. increased $F_{ST}$). Therefore, there was a negative relationship between midparent heterosis and inbreeding depression, the latter expected to be high when panmictic populations are used and low when there is no more heterozygosity in the parents. Hence, having a deeper understanding of inbreeding could lead to better predictions of heterosis. For the 44 ex-PVP lines, long IBD segments (>14.5 Mb) were predominant between ex-PVP and key ancestor lines, suggesting a recent inbreeding, originated in less than 15 generations. There was a high genetic diversity between heterotic groups (stiff stalk and non-stiff stalk), with a reduced diversity among lines in the stiff stalk group. Consequently, we found that a small group of ancestors have contributed large proportions of the genome to important PVP lines in the U.S. Corn Belt. Finally, our results provide a high-resolution data analysis that helps in the identification of IBD regions in the genome that could be used in the quantification of the degree of divergence between parents ($\Delta$), constituting a way to identify the best combination of parents to maximize heterosis.
CHAPTER 1. GENERAL INTRODUCTION

Hybrid-seed production system

G.H. Shull and E.M. East during 1908, independently proved that F1 offspring produced after the crossing of two inbred lines were usually more vigorous and uniform than its parents (East 1908; Shull 1908). However, parental inbred lines developed from open-pollinated varieties were weak with higher susceptibility to biotic and abiotic stresses and often fail to produce good amounts of seed (East 1908; Shull 1908). Therefore, it was not until 1918, after D.F. Jones proposed the use of double-cross hybrids when the exploitation of hybrid vigor starts in the U.S. (Jones 1918). Sprague (1983) and Tracy and Chandler (2006), state that it was in the 1930s when the first double-cross hybrid was first sold in the Midwest and that by the 1960s all of the area planted with corn in the U.S.A corresponded to hybrid seed. The overwhelming success in the use of double-cross hybrids brought the attention of breeders to the production of new inbred lines, resulting in lines that were more vigorous cycle after cycle (Tracy and Chandler 2006; Troyer 1999).

The production of single-cross hybrids from strong inbred parents results in a new era of breeding for corn in the U.S.A., with breeding efforts focusing not only on the production of better inbreds but on the grouping of lines because of their good ability to produce hybrids. As a consequence, inbred lines started to be assigned to heterotic groups, leading to the differentiation of the three most important groups in the U.S. Corn belt: Iowa Stiff Stalk (SSS), Non-Stiff Stalk (NSS), and Iodent (Mikel and Dudley 2006; Tracy and Chandler 2006; van Heerwaarden et al. 2012). By the 1960s and 1970s, several inter-heterotic group crosses were predominant, with the most common crosses being between inbred lines derived from the SSS and NSS heterotic groups (Tracy and Chandler 2006). With time, a consensus was achieved in terms of within
heterotic group breeding schemes, confining the derivation of inbred lines to crosses within groups and to the evaluation of test crosses with lines from opposite groups to select the best performant inbreds (i.e. incorporation of general combining ability). Indeed, this marked a crucial step in the development of new hybrids in the last 80 years, where the production of better and genetically divergent inbred lines has given the key success of the hybrid-seed production system.

Several authors including, Richey (1927), Jenkins (1936), Sprague and Jenkins (1943), Griffing (1954), Sprague (1955), Lonnquist and Gardner (1961), Moll et al. (1962), and Baker (1984), among others, dedicate significant efforts in understanding the basis of the inbred-hybrid system, concluding that hybrid vigor was connected with the levels of genetic diversity in the parents, with the reduction in performance of the parents because of inbreeding (i.e. inbreeding depression), with combining ability, with genetic complementation, with the levels of dominance for the trait, and with the levels of heterozygosity in the offspring (i.e. FST). Moll et al. (1962), concluded that heterosis is dependent on the level of diversity of the parental inbred lines, while Lamkey and Edwards (2004) states that heterosis depends as well on inbreeding depression. According to Hartl et al. (1997), divergent subpopulations can be obtained by isolation, by the lack of genetic flow, and/or by a reduction in the effective population size of each subpopulation, resulting in an increased probability of crossing individuals related by ancestry. Consequently, an increase in the level of homozygosity within subpopulations increases the expression of deleterious recessive alleles (Charlesworth and Willis 2009). After several cycles of within-population crosses, the levels of heterozygosity decreases, reducing the mean phenotypic value when compared to the base population. Therefore, within-heterotic group crosses results in an increased genetic divergence in allele frequency between groups because of the fixation of
alternative alleles and in a decreased heterozygosity, which will decrease the effect of inbreeding depression. When two divergent populations are crossed, complementation between loci carrying deleterious alleles is produced, resulting in an offspring with higher agronomic performance (Lynch 1991). Therefore, decreasing the heterozygosity within a population, an increase in the genetic differentiation between parents and a decrease in the levels of inbreeding depression is obtained. In that regard, the performance in the hybrid could be understood as a function of divergence in allele frequency in the parents, which starts to increase as the population with finite effective population size start to fix different alleles (Wright 1951).

Heterosis

The superiority of the hybrid over its parents by an increase in the mean of crossbred individuals is called heterosis (Bernardo 2010; Falconer and Mackay 1996). This increased performance was originally designated as the result of the contrast product of the union of parental gametes, showing an increase in the level of heterozygosity (Shull 1909, 1952). Lamkey and Edwards (1999) and Kaeppler (2012) states that heterosis is dependent on directional dominance and the square of the genetic divergence in allele frequency between the parents. Guo et al. (2014), working with 294 F8 recombinant inbred lines derived from a single seed descent method, concluded that dominance was the main factor explaining heterosis. Melchinger et al. (2007), found significant dominant effects, concluding that partial to complete dominance (d ≤ 1) are key contributors to heterosis for rosette diameter, leaf area, growth rate per day, biomass yield, and dry matter in Arabidopsis thaliana. Kaeppler (2012), states that heterosis depends not only on the level of dominance but on the genetic divergence in allele frequency between the parents, resulting in differences in genes, pathways, and processes still to be understood and discovered. In that respect, Betrán et al. (2003) and Reif et al. (2003), found significant positive correlations between the genetic distance of the inbred parents and midparent heterosis. Kustanto
et al. (2012) found a correlation ranging from -0.120 for days to pollen shed to 0.181 for grain yield in maize, revealing that by increasing the genetic distance of the parents, an increase in the levels of heterosis was obtained. Tao et al. (2010), using SSR markers, found a correlation between genetic distance and heterosis in rice, concluding that for some traits there was no significant correlation. That goes with Kaeppler (2012) who concludes that genetic distance is not the only factor contributing to heterosis, resulting in a concept that must be defined based on the genotype and trait analyzed. Therefore, heterosis can be thought of as depending on the different possible allele combinations of a particular cross (Falconer and Mackay 1996; Kaeppler 2012).

Heterosis has been defined under three operational classifications. The first was defined as the deviation of the F1 from the mid-parent value and was referred to as midparent heterosis (Bernardo 2010; Falconer and Mackay 1996; Shull 1952). The second was defined as a deviation of the F1 from the best parent and was named high or best parent heterosis (Bernardo 2010). Finally, the third definition was given in the context of commercial production as the superiority of the hybrid compared to the high productive non-hybrid available in a region (Bernardo 2010). The genetic expectation of the second and third definitions depends not only on dominance and genetic divergence in allele frequency but on single-locus additive effects (Melchinger et al. 2007). As reported in the literature, heterosis depends on dominance, the genetic distance between parents, and/or additive epistatic interactions, but not on single-locus additive effects (Bernardo 2010; Falconer and Mackay 1996; Lynch and Walsh 1998), and for that, from the quantitative genetics theory point of view, both definitions fail to provide a good interpretation of the basis of heterosis. However, as pointed out by Melchinger et al. (2007), both definitions can successfully be used to identify, in breeding programs, the hybrid that overcomes the
performance of the better parent and thus, help on the decision-making process in a hybrid-seed breeding system. Therefore, the present dissertation will focus on the interpretation of midparent heterosis (first definition), and in the relationship with genetic divergence in allele frequencies and inbreeding.

**Inbreeding and inbreeding depression**

The term inbreeding has been used in quantitative genetics to denote the mating between individuals with at least one common ancestor (Bernardo 2010; Falconer and Mackay 1996). The degree of inbreeding is measured with the coefficient of inbreeding ($f$), which quantifies the probability that two alleles are identical by descent (IBD) at a given locus in an individual (Bernardo 2010; Falconer and Mackay 1996; Malécot 1948; Rudan and Campbell 2004). Under an idealized population, the probability that one allele will unite at random with an identical allele is twice the population size ($2N$) (Falconer and Mackay 1996). When considering the inbreeding for future generations, $f$ is partitioned in an estimation of the increment attributable to new inbreeding ($2N^{-1}$) and an increment attributable to old inbreeding given by $(1-(2N^{-1})f_{t-1}$, where $t$ refers to the current generation (Falconer and Mackay 1996). The levels of inbreeding will vary depending on how closely related the individuals are, how different the contribution of males and females is, how unequal the distribution of family sizes from generation to generation is (effective population size), and how co-selection of relatives works in the population (Falconer and Mackay 1996).

The immediate effect of inbreeding is a significant reduction in the mean phenotypic value of a population compared to a reference population because of the degeneration of fitness, growth rates, and/or an increase in the susceptibility to diseases and environmental stresses, among others (Charlesworth and Willis 2009; Falconer and Mackay 1996; Goff 2011; Jones 1917). This reduction, known as inbreeding depression, is often caused by increased
homozygosity, rising the probability of alleles identical by descent (Bernardo 2010; Charlesworth and Charlesworth 2009; Charlesworth and Willis 2009; Jacquard 1975; Lynch and Walsh 1998; Rudan and Campbell 2004). Changes in the population mean caused by inbreeding are associated with the dominance effect of a locus ($d > 0$), while the direction of the change will depend on the allele with the most recessive effect (Charlesworth et al. 1987; Falconer and Mackay 1996). Additionally, the magnitude in which the reduction in the mean is achieved will depend on allele frequencies, being maximum when they are 0.5, and absent when the allele becomes fixed in the population (Falconer and Mackay 1996). Several studies have indicated the importance of inbreeding depression in several species. For instance, Jain and Bharadwaj (2014), reported in maize, a maximum reduction in the mean of the $F_2$ generation compared to the $F_1$ (i.e. inbreeding depression) of 12.42% for shelling percentage, 10.34% for protein percentage, and 18.58% for lysine content. Goff (2011), reported that most inbred lines of maize suffer from yield losses between 40 to 60 percent, or even more when compared to crossbred individuals. Equally, Edwards and Lamkey (2002), reported rates of inbreeding depression ranging from 0.01 to 0.045 Mg/ha for grain yield, 0.3 to 0.6 cm for plant height, and -0.04 to -0.07 days for anthesis once the percentage in the coefficient of inbreeding increases.

Heterosis and inbreeding depression are known to be opposite phenomena that depend on allele frequencies and directional dominance. Differences in performance between hybrids and inbreds have been attributed to changes in gene expression and metabolic rates (Goff 2011; Paige 2010). According to Goff (2011), this difference leads to a deficiency in the production of proteins and/or enzymes in inbred individuals because of different enzymatic or transcriptomic pathways. Also, homozygosity caused by inbreeding is related to an increase in energy consumption and oxygen requirements, affecting genes associated with basic metabolism
(energy use), stress tolerance, cellular maintenance, and defense mechanisms (Ketola and Kotiaho 2012; Kristensen et al. 2005; Paige 2010), with fitness-related traits more sensitive to inbreeding than morphological traits (Ketola and Kotiaho 2009).

Quantitative genetic studies of heterosis and inbreeding

Willham and Pollak (1985) developed a theory for midparent heterosis in which the random-mated cross of the F1 between two populations, denoted as F2, was the reference population. They express heterosis for a one-locus two alleles case, as the result of the square of the difference in allele frequencies between the parents (Δ^2) and directional dominance (d), while for a case of 2 alleles 2 loci, heterosis has additionally an additive-by-additive epistatic effect, which if positive will reduce the amount of heterosis expressed in the F1. In the same sense, Hill (1982), defined a model in which the F2 generation, produced after crossing two inbred lines, was used as the reference population. Hill (1982), uses the parameterization proposed by Cockerham (1954) and Kempthorne (1954), in which the genetic composite effects were described by the contribution of additive (a), dominance (d), and/or epistatic effects. In doing so, Hill (1982), expressed heterosis by the contribution of dominance and additive-by-additive epistatic effects, while the genetic mean for each generation was defined as the combination of the genic divergence in allele frequency between the two inbred parents and the gene effects. Lynch (1991) uses Hill’s derivations to develop a framework to understand the genetic bases of inbreeding and outbreeding depression. He described heterosis by using two summary coefficients, a source index (θS), which relates the alleles in the offspring with the parent of origin, and a hybridity index (θH), which provides information about the inbred/crossbred status of an individual. As Willham and Pollak (1985) and Hill (1982), Lynch (1991) results suggest that the enhancement observed in the F1 generation because of outcrossing (heterosis) could be
explained by the dominance effect of favorable alleles, by additive-by-additive epistatic interactions, or by both effects.

Lynch (1991), additionally, considered a scenario in which the parent populations were inbred to the level of inbreeding \( f \), however, the complexity of the model increases, yielding a total of fifteen unknown parameters to be estimated. Lamkey and Edwards (1999) extended Willham and Pollak’s (1985) theory to include any level of inbreeding in any generation, allowing all individuals within a generation to have loci with an \( f \) probability of alleles identical by descent. In their estimation, the mean of each generation constitutes a weighted average of \( 1-f \) times the mean at panmixia plus \( f \) times the mean at complete inbreeding. Lamkey and Edwards (1999) model expressed the genetic mean for each generation as the combination of the genetic divergence in allele frequency (\( \Delta \)), the individual gene effects (\( a \) and \( d \)), the average effect of an allele substitution (\( \alpha \)), and the level of inbreeding \( f \), demanding at least 5 generations to obtain estimates for each of the genetic parameters. In conclusion, heterosis at a single-locus 2 alleles case has been described to be dependent on dominance, the genetic divergence in allele frequency between the parents, and inbreeding depression, while when a multi-locus and/or multiple alleles case is assumed, the contribution of additive-by-additive and dominant-by-dominant epistatic effects can increase or reduce the level of heterosis shown in the crossbred generation.

**Detection of identity-by-descent (IBD) segments with genome-wide genotypic data**

All gene copies present in an individual could be the result of DNA replication of one of the \( 2N \) genes present at a locus in a base population (Hartl et al. 1997; Walsh and Lynch 2018). These genes or alleles are referred to as being identical copies coming from a common ancestor or simply alleles identical-by-descent (Hartl et al. 1997; Malécot 1948; Wright 1921). Sved (1971), states that genes inherited from a common ancestor (IBD genes) are arranged in
chromosome segments that have a lack or reduced rate of crossing over with homologous segments. Thompson (2013), suggests that large DNA segments containing several IBD alleles are inherited from generation to generation, with a decreasing length once the number of recombination events or linkage disequilibrium increases, or when the effective population size decreases. Therefore, it has been proposed that the length of an IBD segment follows an exponential distribution with parameter λ equal to twice the number of meiosis events separating the individuals (Fisher 1954; Stam 1980; Thompson 2013). Bjelland et al. (2013) state that long segments with consecutive homozygote markers are associated with inbreeding in recent generations during which recombination has not been able to break up the segment. Consequently, as the number of meiosis separating an individual with the ancestor increases, the expected length of the IBD segment decreases, reducing at the same time the ability to correctly identify them.

The first and simplest approach to account for IBD segments from genotypic data consisted of counting in a pair of individuals the number of identical markers, followed by calling an IBD segment if the length of the segment overcome a fixed threshold. Even when this process constitutes an easy and fast way, it is not able to distinguish between segments that are IBD and segments that are identical by state (Bjelland et al. 2013). As stated in the literature, DNA segments that are identical by descent are, by definition, identical by state (IBS), but not all IBS segments are copies coming from a common ancestor (Bernardo 2010; Falconer and Mackay 1996; Hartl et al. 1997; Walsh and Lynch 2018), and thus, this approach gives an overestimated value of IBD segments between a pair of individuals. With that in mind, two major methods were proposed to account for IBD segments using genotypic data. The first corresponds to an observational genotype or haplotype counting algorithm that looks for matches
along each chromosome by moving a user-defined sliding window (Browning and Browning 2012; Ceballos et al. 2018). The sliding windows work by splitting the genome based on the number of SNPs specified (user-defined) and moving along the chromosome one SNP at a time (Bjelland et al. 2013). A ratio between the number of completely homozygote windows over the total number of windows in which an SNP was included is computed for all sets of SNPs in the covered genome (Bjelland et al. 2013; Browning and Browning 2012). Finally, the programs called an IBD segment based on the number of consecutive homozygote SNPs and/or the length of the segment exceeding a user-defined threshold (Browning and Browning 2012; Ceballos et al. 2018).

The second methodology corresponds to a model-based probabilistic approach using hidden Markov models (HMM) in the estimation of the IBD status at each locus (Browning and Browning 2012; Ceballos et al. 2018). This type of model works as a normal Markov chain in which the internal states, here the IBD status of each SNP, is not known and can only be estimated through the use of a prior, emission, and transitions probabilities. The prior probability describes the probability of IBD for the first marker, while the transition probability describes the likelihood of transitioning from the IBD status of maker $k-1$ to maker $k$ (Han and Abney 2011; Leutenegger et al. 2003; Westhead and Vijayabaskar 2017). Finally, the emission probability describes the likelihood of maker $k$ being homozygote or heterozygote given the unknown hidden state of that marker (Westhead and Vijayabaskar 2017). IBD transition probability depends on the distance between marker $k-1$ and $k$, and thus variation in the rate of recombination is included in the model (Browning and Browning 2012; Leutenegger et al. 2003). When high-density SNP data sets are used, additional care about linkage and linkage disequilibrium (LD) between closely spaced markers must be considered. When the model does
not consider LD in the computation, an increase in the identification of false IBD segments is produced (Browning and Browning 2012). Several authors have suggested that a thinning of markers in LD before running such a model will result in a decrease of false IBD segments, providing an accurate IBD probability for each marker (Browning and Browning 2007; Browning and Browning 2012; Chang et al. 2015; Huff et al. 2011; Pemberton et al. 2012; Purcell et al. 2007), however, this reduces the total amount of markers that could be included in the analysis. The incorporation of LD in the HMM framework has resulted in a complex model that has a modified emission probability to account for LD, resulting in a conditional likelihood of the genotype at a marker k, given the IBD status and the genotype of m adjacent loci (Browning and Browning 2012; Han and Abney 2011, 2013). By doing so, the rate of false IBD segments is decreased significantly at the same time of using all SNP markers available in the study.

**Objectives**

The main objective of this dissertation was to understand the underlying basis of heterosis, to estimate the genetic diversity and progenitor’s genetic contribution based on the amount of shared IBD segments, and to understand the statistical models used in genome-wide association studies. Specifically, the following areas will be discussed: 1) to test a model of heterosis based on genetic divergence in allele frequency between the parents (Δ), dominance, and inbreeding depression, 2) to estimate the contribution of Δ and the inbreeding depression in the parents to heterosis, and 3) to estimate the genetic importance of a hybrid seed production system for grain yield, plant height, and ear height, 4) to quantify the size of IBD segments, 5) to estimate the number of generations from a common ancestor for each segment, 6) to quantify the genetic diversity of with 44 ex-PVP and eight key ancestor lines when computed from IBD segments, 7) to compared diversity estimates calculated from pedigree and IBD segments, 8) to
quantify the genetic contribution of the eight key ancestors, and 9) to describe the most common statistical models used in genome-wide association studies (GWAS).

**Dissertation organization**

This dissertation is organized into five chapters. Chapter 1 is the introduction, which provides an explicit explanation of the objectives and literature review, Chapters 2, 3, and 4 are manuscripts to be submitted to Theoretical and Applied Genetics (TAG), and Chapter 5 contains the general conclusions. Since Chapters 2, 3, and 4 are manuscripts being submitted to TAG, the entire dissertation will follow the style of literature citations used in TAG. The manuscript contained in Chapter 2 will address objectives 1 - 3, the manuscript in Chapter 3 will cover objectives 4 and 8, and the manuscript in Chapter 4 will address objective 9. In a more general sense, Chapter 2 describes the underlying basis of heterosis, Chapter 3 the use of inbreeding and IBD segments to estimate genetic diversity and progenitors’ genetic contribution, and Chapter 4 to the principles behind some common methods used in genome-wide association studies (GWAS).
CHAPTER 2. POPULATION DIVERGENCE AND INBREEDING DEPRESSION IN MAIZE HETEROSIS

Aguilar, Fernando S.¹, Lamkey, Kendall R.¹, Edwards, Jode W.²

¹ Department of Agronomy, Iowa State University, Ames, IA 50011, USA
² USDA-ARS Corn Insects and Crop Genetics Research Unit, Ames, IA 50011, USAAA

Modified from a manuscript to be submitted to Theoretical and Applied Genetics

Abstract

Heterosis refers to the superiority of the hybrid over its parents. Heterosis depends on non-additive gene effects and in the difference in allele frequencies between parents. The objective of this experiment was i) to test a model of heterosis based on dominance, genetic divergence in allele frequency between the parents (Δ), and inbreeding depression, ii) to estimate the contribution of Δ and the inbreeding depression in the parents to heterosis, and iii) to estimate the genetic importance of a hybrid seed production system for grain yield, plant height, and ear height. Our results showed that a single-locus model with two alleles explains the variability in each generation and was suitable to predict heterosis in maize. Midparent heterosis (MH) is a function of Δ and dominance, while inbred-midparent heterosis depends on Δ, dominance, and inbreeding depression. There was a negative relationship between MH and inbreeding depression. Inbreeding depression is expected to be high in panmictic populations, however, when $F_{ST}$ approaches unity inbreeding depression is reduced to zero. Lower midparent values and higher MH were obtained when the parents were inbred to homozygosity, with a limiting value being an inbred line. Reduced heterozygosity, which results in low inbreeding depression, could be understood as the underlying basis of heterosis. Consequently, the recovery from inbreeding depression because of the masking of recessive alleles could result in high heterosis.
The genetic value of a hybrid seed production system can be attributed to half midparent heterosis, allowing breeding schemes to fix and maintain the best genotypes.

**Introduction**

George Shull (1914) first proposed the term heterosis to refer to the superiority of the hybrid over its parents reflected by an increase in the mean trait values of crossbred individuals for grain yield, plant height, among other traits (Charlesworth and Willis 2009; Falconer and Mackay 1996; Shull 1914). Heterosis is a function of non-additive gene action and the square of the difference in allele frequencies between the parents (Falconer and Mackay 1996; Lamkey and Edwards 1999). According to Kaeppler (2012), heterosis depends on the genetic divergence in allele frequency between parents, resulting in differences in genes, pathways, and processes still to be understood and discovered. Marcón et al. (2019) found that by increasing the genetic distance between the parents in *Paspalum notatum*, the proportion of aposporous hybrids increases, while Reif et al. (2003) found a significant positive correlation ($r = 0.63$) between genetic distance and midparent heterosis for grain yield in maize. Tao et al. (2010), studied, with SSR markers, the correlation between the genetic distance of parents and heterosis in rice, finding that for yield there was a significant positive correlation but for number of panicles per plant, number of grains per panicle, and 1000-grain weight there were no significant correlations. That goes with Kaeppler (2012) who concluded that the genetic divergence between the parents is not the only factor contributing to heterosis and that other factors can be influencing the expression of hybrid vigor. As a result, heterosis has been described to be genotype and trait dependent, conditioned on the different possible allele combinations of a particular cross (Falconer and Mackay 1996; Kaeppler 2012), and in the degree of the genetic difference between the parents (Tao et al. 2010).
An increase in the genetic difference in allele frequency between populations increases the relative between-populations heterosis and decreases the within-population inbreeding depression (Charlesworth 2018). Charlesworth and Willis (2009), mention that lower fitness within a population could arise from increased homozygosity of recessive deleterious alleles and/or alleles with a heterozygote advantage. The consequence of inbreeding is characterized by an increased expression of recessive deleterious alleles, which are present in a population at low frequencies (Charlesworth and Willis 2009; Wang et al. 1999). Latter (1998) found that several small-effect deleterious mutations were responsible for the loss of fitness in the population and the lower rate of inbreeding in *Drosophila*. Similarly, Lariepe et al. (2012) indicate that maize could have accumulated more deleterious alleles than rice (an autogamous plant), because these alleles could have been masked by the effect of normal alleles in its homologous chromosome. Because of their recessive nature, these deleterious alleles are masked in heterozygote individuals through balancing selection, chromosomal inversion, and/or polymorphisms for regions of the genome with suppressed recombination (Charlesworth and Willis 2009; Paige 2010; Wang et al. 1999). Then, a significant reduction in the mean phenotypic value compared to the base population (i.e. inbreeding depression) is obtained by the degeneration of fitness, growth rates, basic metabolism, and an increased energy consumption and oxygen requirements, among others (Charlesworth and Willis 2009; Falconer and Mackay 1996; Goff 2011; Jones 1917; Kristensen et al. 2005; Myrand et al. 2002). Inbreeding depression and heterosis have been referred to as opposite phenomena that depend on allele and genotype frequencies. Heterosis is the product of interacting genomes resulting from the combination of alleles from different sources, with a higher impact when the parents are genetically divergent (Chen 2013; Goff 2011). An increase in the level of genetic divergence is achieved when the level of homozygosity
increases (Kristensen et al. 2006). However, once the population approaches complete inbreeding ($F_{ST} = 1$), the decrease in the phenotypic mean because of inbreeding is significantly reduced. For instance, Wang et al. (1999), found that after several cycles of inbreeding the decrease in fitness in the population becomes slow, and the difference in allele frequency between populations increases.

Willham and Pollak (1985) developed a theory to describe the mean performance of parents and crosses by considering the $F_2$ generation as the reference population from which genetic inferences were made. They express heterosis for a one-locus with two alleles case to be proportional to the dominance effect ($d$) and the square of the difference in allele frequencies between the parents ($\Delta^2$). When expanding the theory to include two loci, Willham and Pollak (1985) found that heterosis has additionally an additive-by-additive epistatic effect, which could reduce or increase the amount of heterosis compared to cases in which epistasis is absent or negligible. Lynch (1991), using the same assumptions, described the mean of each generation in terms of a source index ($\theta_S$), which relates the alleles in the offspring with the parent from where they were inherited, and a hybridity index ($\theta_H$), which provides information about the inbred/crossbred status of an individual. Both Willham and Pollak (1985) and Lynch (1991), concluded that the enhancement observed in the $F_1$ generation because of outcrossing could be explained by the dominance of favorable alleles present in the parent populations, by favorable additive-by-additive epistatic interactions, or by both effects. Willham and Pollak (1985) and Lynch (1991) theories highlight the importance of divergence in allele frequencies between the parents ($\Delta$), concluding that all pairs of genes with non-additive effects from divergent parents have the potential to enhance the performance of the $F_1$. Lynch (1991) developed some models for cases in which the parent population can be inbred to an $f$ level, showing that dominance,
additive-by-dominance, and dominance-by-dominance effects were confounded. However, by including both the inbreeding and crossbreeding effect to describe the mean of the generations, Lynch (1991) developed a complex formula that includes fifteen unknowns, forcing the use of more than fifteen types of generations to estimate the parameters in the model.

Lynch (1991) stated that the genetic divergence between the parents had a great influence in the expression of grain yield heterosis, but that the greatest advantage was observed at intermediate levels of genetic distance. According to Lamkey and Edwards (1998), the level of dominance and genetic divergence in allele frequency between parents are the main effects for the expression of heterosis in the F1 produced from panmictic parents, while heterosis in the cross of inbred lines has the contribution of dominance, inbreeding depression, and genetic divergence. Lamkey and Edwards (1998) concluded that most of the heterosis observed in the F1 between inbred lines was caused by the recovery in performance from what was lost during inbreeding and that sometimes genetic divergence could have little effect on heterosis. The recovery in performance from inbreeding depression has been referred to as baseline heterosis and represents the average amount of heterosis that can be obtained from crossing all possible inbred lines derived from two different panmictic populations (Lamkey and Edwards 1998; Lamkey and Edwards 1999). Therefore, baseline heterosis quantifies the average inbreeding depression observed in the parent populations (Lamkey and Edwards 1999). In the present study, we refer to baseline heterosis as the average inbreeding depression of the parents because baseline heterosis is an estimate of inbreeding depression in the parents. Therefore, it is our objective i) to test a model of heterosis based on dominance, genetic divergence in allele frequency between the parents (Δ), and inbreeding depression, ii) to estimate the contribution of
Δ and the inbreeding depression in the parents to heterosis, and iii) to estimate the genetic importance of a hybrid seed production system for grain yield, plant height, and ear height.

Materials and methods

Genetic materials

Six synthetic maize populations and eight inbred lines were used in the experiment (Table 2.1). All populations were derived from 3 different synthetics, BSSS, BSCB1, and BS11. The synthetic BS13(S)C10 is a direct descendent of the BSSS synthetic population (Table 2.1). Both B129 and B73 inbred lines are members of the stiff stalk heterotic group, while the remaining inbred lines are members of the non-stiff stalk heterotic group. Three crosses were made between synthetic populations, three between synthetic populations and the B129 inbred line, and six between inbred lines (Table 2.2). Fourteen parents, corresponding to eight inbred lines and six synthetic populations were used in the experiment (Table 2.1). For each of the twelve crosses listed in table 2.2, seed was produced for parents (P1 and P2), F1 hybrid, self-pollinated F1 (F1-selfed), random mated F1 (F2), and backcrosses to parent 1 (BCP1) and 2 (BCP2) for a total of 60 pedigrees derived from crosses. Each of the six synthetic parent populations was self-pollinated for one generation to produce the S1 generation. Thus, a total of 80 pedigrees corresponding to 14 parents, 6 S1, and 60 cross-derived pedigrees, were used in the experiment.

Experimental design

The 80 pedigrees were grown in five locations near Ames, Carroll, Crawfordsville, Fairfield, and Lewis, Iowa during the growing seasons of 2007 and 2008. Sixteen of the 80 unique pedigrees were replicated multiple times within each replicate-block in each location resulting in 100 experimental units per replicate. The experiment was arranged in a modified split-plot design with three replications where inbreeding level (F1, F2, Backcrosses, Inbred lines, and synthetic Populations) was considered as the whole-plot treatment factor and pedigree within
inbreeding level as the subplot treatment factor. The design differed from a classical split-plot because subplots (pedigrees) were not cross-classified with whole plots and because each level of the whole plot factor was applied to more than one whole-plot experimental unit in a replicate. The experimental unit for the subplot factor was a plot of four rows, spaced 0.76 cm apart and 5.49 m long. The whole-plot experimental unit was a block containing five subplots side-by-side (20 rows, 5.49 m long), with one of five possible inbreeding levels applied to each whole-plot experimental unit. Each replicate containing 100 subplots was divided into 20 whole plot blocks which were then separated into 2, 3, 3, 5, and 7 blocks for the synthetic populations, inbred lines, F1, backcrosses, and F2 (including the selfed F1) inbreeding levels, respectively. Each whole plot experimental unit was randomly assigned to a range in the field. Within each replicate-block, there were 10, 15, 15, 25, and 35 subplots (pedigrees) for synthetic populations, inbred lines, F1, backcrosses, and F2, respectively. The seed was sown at a density of 7.0 plants per m² using an Almaco mini-belt cone planter. Agronomic practices were applied to each experiment following commercial maize production practices in central Iowa. The sampling unit consisted of the two central rows of each plot to avoid border effects.

**Data collection**

Data was collected on a plot basis in the two center rows for plant height (cm), ear height (cm), and grain yield (Mg ha⁻¹) adjusted to 15.5% grain moisture. Plant height (cm) and ear height (cm) were recorded during 2007 and 2008 for the experiments grown near Carroll and Lewis, and during 2007 for Crawfordsville. Plant height was measured by taking the distance from the soil surface to the flag leaf collar. Similarly, ear height was measured as the distance from the soil surface to the uppermost ear node. The mean value for plant and ear height was recorded as the average of ten randomly selected plants per plot. Plots were harvested with a New Holland TR88 combine modified for automatic acquisition of test weight, grain moisture,
and grain weight. Accurate grain moisture measurement requires a minimum of approximately 1.8 kg. The location Fairfield, 2007, had 20% of plots with grain mass less than 1.8 kg and thus, was dropped from the analysis for grain yield.

**Data analysis**

Data were analyzed by fitting the linear mixed-effects model:

\[
Y_{ijklmn} = \xi_i + \Gamma_{l(i)} + I_j + \omega_{ij} + \psi_{j(k)} + \phi_{ik(j)} + \nu_{m(l)} + \delta_{n(i)} + \varepsilon_{ijklmn}
\]

Where \(Y_{ijklmn}\) was the response variable for range \(n\), pass \(m\), replication \(l\), pedigree \(k\), inbreeding level \(j\), and environment \(i\), \(\xi_i\) the effect of environment \(i\), \(\Gamma_{l(i)}\) the replication \(l\) within environment \(i\), \(I_j\) the inbreeding level \(j\), \(\omega_{ij}\) the interaction between environment \(i\) and inbreeding level \(j\), \(\psi_{j(k)}\) the pedigree \(k\) within inbreeding level \(j\), \(\phi_{ik(j)}\) the interaction between the pedigree \(k\) within inbreeding level \(j\) and environment \(i\), \(\nu_{m(l)}\) the pass \(m\) within replication \(l\) and environment \(i\), \(\delta_{n(i)}\) the range \(n\) within replication \(l\) and environment \(i\), and \(\varepsilon_{ijklmn}\) the residual error. All effects were fitted as fixed effects except the environmental interactions \(\omega_{ij}\) and \(\phi_{ik(j)}\), and the range within replication and environment, \(\delta_{n(i)}\). Outliers were identified by fitting the full linear model and estimating the probability of obtaining a larger absolute value for each residual using a t-distribution and adjusting individual p-values with a Bonferroni correction at a 2% level of significance. After removing outliers, variances of residuals \(\varepsilon_{ijklmn}\) and the range within replication and environment \(\delta_{n(i)}\) were considered to be heterogeneous among environments such that \(V(\varepsilon_{ijklmn}) = \sigma^2_{e(i)}\) and \(V(\delta_{n(i)}) = \sigma^2_{\delta(i)}\). Twenty-four models, produced after fitting all possible combinations in both the random effects and the homogeneity/heterogeneity in the residual variance \(V(\varepsilon_{ijklmn}) = \sigma^2_{e} \quad \text{vs.} \quad V(\varepsilon_{ijklmn}) = \sigma^2_{e(i)}\) and the range within replication and environment variance \(V(\delta_{n(il)}) = \sigma^2_{\delta} \quad \text{vs.} \quad V(\delta_{n(il)}) = \sigma^2_{\delta(i)}\), were analyzed (Table 2.4). Bayesian Information Criterion (BIC) was used to identify the best fitting model (Schwarz 1978). Residual
Maximum Likelihood (ReML) was used to obtain the estimates of the variances by using the Fisher-Scoring algorithm in ASReml (Gilmour et al. 2015). These estimates were used to fit a mixed model in SAS software (SAS Institute, Cary, NC) to compute tests of fixed effects and best linear unbiased estimators (BLUE). Least squares mean was obtained for each environment, inbreeding level, and pedigree within inbreeding level, and used to estimate heterosis and inbreeding depression. All analyses were done using the MIXED procedure of SAS 9.4 software (SAS Institute, Cary, NC) and ASReml release 4.1 software (Gilmour et al. 2015).

**Genetic model**

Consider an ancient population (metapopulation) created some time in the past from where several subpopulations were derived. These subpopulations, with finite effective population size \((N_e)\), are formed by a group of genotypes selected from the metapopulation. Therefore, the average inbreeding of each subpopulation relative to the metapopulation will be bigger than zero \((F_{ST} > 0)\), resulting in a divergence in allele frequency between subpopulations bigger than zero \((\Delta > 0)\). Subsequently, any change in \(N_e\) will directly result in an increase or decrease in \(F_{ST}\) and thus in \(\Delta\). Willham and Pollak (1985) developed a theory for midparent heterosis where the random-mated cross between two such subpopulations, denoted as F2 generation, was the reference population (Figure 2.1). They consider a simple one-locus model where \(p_i\) and \(p'_i\) were the allele frequencies for allele \(A_i\) in parents 1 and 2, respectively. In their model, Willham and Pollak (1985) expressed the allele frequencies for \(A_i\) and \(A_j\) as linear combinations of the average \((\bar{p}_i = \frac{1}{2}(p_i+p'_i))\) and differences \((\delta_i = \frac{1}{2}(p_i - p'_i)\) or \(\delta_j = \frac{1}{2}(p_j - p'_j))\) in gene frequencies between the two populations, such that the gametic array could be expressed as \(p_i = \bar{p}_i + \delta_i\) for parent 1 (P1) and \(p'_i = \bar{p}_i - \delta_i\) for parent 2 (P2). Under this conditions, the F2 generation, produced by random mating the F1, constitutes a population in
Hardy Weinberg equilibrium from where the genotypic array can be expressed as the product of the gametic array at each locus in the original population (Kempthorne 1957), such that:

\[ F_2 = \sum_{ij} \bar{p}_i \bar{p}_j Y_{ij} \]

Where \( Y_{ij} \) corresponds to the genotypic value associated with genotypes \( A_iA_i, A_iA_j, \) and \( A_jA_j \). In quantitative genetics theory, half the difference between the two homozygotes is defined as the additive effect \( a \), while the difference between the heterozygote and the mean of the two homozygotes is referred to as the dominant effect \( d \) (Bernardo 2010; Falconer and Mackay 1996). Therefore, the genotypic value associated with the homozygotes is \( a \) for \( A_iA_i \) and \(-a\) for \( A_jA_j \) (assuming \( A_i \) is the allele that increases the mean), while it is \( d \) for the heterozygotes (Bernardo 2010; Falconer and Mackay 1996; Willham and Pollak 1985). Thus, the genotypic array of the \( F_2 \) generation is a function only of average allele frequency and dominance and for that, it is logical to use it as the reference population. The mean of the generations of interest (\( P_1, P_2, F_1, \) and backcrosses to either parent), expressed as the product of the gametic array at each locus and the genotypic value \( (Y_{ij}) \), are:

\[ P_1 = \sum_{ij} (\bar{p}_i + \delta_i)(\bar{p}_j + \delta_j) Y_{ij} = \sum_{ij} \bar{p}_i \bar{p}_j Y_{ij} + \sum_{ij} \delta_i \delta_j Y_{ij} \]

\[ P_2 = \sum_{ij} (\bar{p}_i - \delta_i)(\bar{p}_j - \delta_j) Y_{ij} = \sum_{ij} \bar{p}_i \bar{p}_j Y_{ij} - \sum_{ij} \delta_i \delta_j Y_{ij} \]

\[ F_1 (P_1 \times P_2) = \sum_{ij} (\bar{p}_i + \delta_i)(\bar{p}_j - \delta_j) Y_{ij} = \sum_{ij} \bar{p}_i \bar{p}_j Y_{ij} - \sum_{ij} \delta_i \delta_j Y_{ij} \]

\[ BC_{P1} = \sum_{ij} (\bar{p}_i + \delta_i) \bar{p}_j Y_{ij} = \sum_{ij} \bar{p}_i \bar{p}_j Y_{ij} + \sum_{ij} \bar{p}_j \delta_i Y_{ij} \]

\[ BC_{P2} = \sum_{ij} \bar{p}_i (\bar{p}_j - \delta_j) Y_{ij} = \sum_{ij} \bar{p}_i \bar{p}_j Y_{ij} - \sum_{ij} \bar{p}_i \delta_j Y_{ij} \]

For easy deductions, the model was reduced to a case with two-alleles in which \( p_1 \) and \( p_1' \) were the gene frequencies of allele 1 in parents 1 and 2, respectively. Defining \( \Delta \) as half the
difference in allele frequencies between the two parents \(\frac{1}{2}(p_i - p_i')\), the terms \(\delta_1\) and \(\delta_2\) reduces to \(\Delta\) and \(-\Delta\), respectively (Willham and Pollak 1985). The terms \(\sum_{i,j} \bar{p}_i \delta Y_{ij}\) and \(\sum_{i,j} \bar{p}_j \delta Y_{ij}\) can be reduced each to \(\Delta[\bar{p}_1(Y_{11} - Y_{21}) + \bar{p}_2(Y_{12} - Y_{22})]\), where the terms within square brackets correspond to the average effect of an allele substitution \(\alpha\) (Willham and Pollak 1985), defined as the change in the genotypic value when one allele is randomly substituted for the other allele (Bernardo 2010; Lynch and Walsh 1998). Similarly, the term \(-\sum_{i,j} \delta_i \delta_j Y_{ij}\) can be expressed as \(\Delta_2(Y_{12} + Y_{21} - Y_{11} - Y_{22})\), which is equivalent to \(2\Delta_2 d\) (Willham and Pollak 1985). Generation means are:

\[
\begin{align*}
F_2 &= \mu \\
P_1 &= \mu + 2\Delta\alpha - 2\Delta^2d \\
P_2 &= \mu - 2\Delta\alpha - 2\Delta^2d \\
F_1 (P_1 \times P_2) &= \mu + 2\Delta^2d \\
BC_{P1} &= \mu + \Delta\alpha \\
BC_{P2} &= \mu - \Delta\alpha
\end{align*}
\]

Where \(\mu\) corresponds to the population mean and can be expressed as \(a(\bar{p}_1 - \bar{p}_2) + 2\bar{p}_1\bar{p}_2d\).

Willham and Pollak (1985) developed their theory in the context of animal breeding where individual inbreeding within populations is not common. Lamkey and Edwards (1999) considered practices in plant breeding in which the development of inbred lines is common and extended Willham and Pollak’s theory to include any level of inbreeding \(f\) in any generation.

Under Lamkey and Edwards (1999) model, the means of the generations are expressed as a weighted average of \(1-f\) times the mean at panmixia plus \(f\) times the mean at complete inbreeding. When \(f\) equals 1, there is no contribution of the dominant effect \(d = 0\) and thus the average effect of an allele substitution \((\alpha = a + d(\bar{p}_2 - \bar{p}_1))\) becomes a function of only the
difference between homozygotes, $\alpha (\alpha = a)$. Correspondingly, the mean of all generations expanded to include any level of inbreeding are:

$$F_2 = (1-f)\mu + fa(p_1 - p_2) = \mu - 2\bar{p}_1\bar{p}_2 df = \mu_f$$

$$P_1 = (1-f)(\mu + 2\Delta \alpha - 2\Delta^2 d) + fa(\bar{p}_1 - \bar{p}_2 + 2\Delta) = \mu_f - 2\Delta^2 d(1-f) + 2\Delta \alpha_f$$

$$P_2 = (1-f)(\mu - 2\Delta \alpha - 2\Delta^2 d) + fa(\bar{p}_1 - \bar{p}_2 - 2\Delta) = \mu_f - 2\Delta^2 d(1-f) - 2\Delta \alpha_f$$

$$F_1 = (1-f)(\mu + 2\Delta^2 d) + fa(\bar{p}_1 - \bar{p}_2) = \mu_f + 2\Delta^2 d(1-f)$$

$$BC_{P1} = (1-f)(\mu + \Delta \alpha) + fa(\bar{p}_1 - \bar{p}_2 + \Delta) = \mu_f + \Delta \alpha_f$$

$$BC_{P2} = (1-f)(\mu - \Delta \alpha) + fa(\bar{p}_1 - \bar{p}_2 - \Delta) = \mu_f - \Delta \alpha_f$$

Where $\mu_f$ corresponds to the mean of the population at inbreeding coefficient $f$ and $\alpha_f$ is the average effect of an allele substitution at inbreeding $f$. The term $\mu_f$ represents a weighted average of the population mean ($\mu$) at panmixia minus the change of mean because of inbreeding $2\bar{p}_1\bar{p}_2 df$, while, $\alpha_f$ represents a weighted average of the coefficient of inbreeding $f$ times the gene substitution at complete inbreeding ($a$) plus $1-f$ the average effect of an allele substitution at panmixia ($\alpha$). In this experiment, Lamkey and Edwards’ (1999) theory was extended to not only include panmictic populations as parents but to include inbred lines in the cross. For inbred lines, the estimated allele frequency at a locus is either zero or one, resulting in values for the average allele frequency and the difference in allele frequency of one half ($\bar{p}_1 = \bar{p}_2 = \Delta = \frac{1}{2}$).

Consequently, the average effect of an allele substitution ($\alpha$) is reduced to the homozygote contrast, $a$. By substituting these values in the above equations, the mean for each generation for inbred line crosses are:

$$F_2 = \mu_f = \mu = \frac{1}{2}d$$

$$\text{Inbred}_1 = \mu_f - 2\Delta^2 d(1-f) + 2\Delta \alpha_f = \mu - 2\bar{p}_1\bar{p}_2 d + 2\Delta a = a$$
Inbred$_2 = \mu_f - 2\Delta^2d(1-f) - 2\Delta \alpha_f = \mu - 2\bar{p}_1 \bar{p}_2d - 2\Delta a = -a$

$F_1 = \mu_f + 2\Delta^2d(1-f) = \mu + 2\Delta^2d = d$

$BC_{\text{Inbred}1} = \mu_f + \Delta \alpha_f = \mu + \Delta a = \frac{1}{2}d + \frac{1}{2}a$

$BC_{\text{Inbred}2} = \mu_f - \Delta \alpha_f = \mu - \Delta a = \frac{1}{2}d - \frac{1}{2}a$

**Heterosis**

The difference in the performance of the crossbred generation ($F_1$) and the average of the parents are defined as midparent heterosis (Falconer and Mackay 1996). In the present, midparent heterosis (MH) was expanded by deviating the $F_2$ generation from the midparent value and it was referred to as midparent $F_2$ heterosis ($MF_2H$) (Table 2.3). Additionally, the difference between the crossbred generation and the mean of its parents when inbred to homozygosity (inbred midparent value, IMPV) was estimated (Table 2.3). For our purposes, we will call this difference as inbred-midparent heterosis (IMH). Inbred midparent values were estimated by selfing each parent population for one generation to create the $S_1$. Then, the difference between two times the mean of the $S_1$ of the parent and the mean of the parent before inbreeding ($2S_{1,q} - P_q$, for $q = 1, 2$) gives origin to the mean of the parent when inbred to homozygosity ($P_{\gamma(f=1), q = 1, 2}$). As before, we deviate the $F_2$ generation from the inbred midparent value to estimate the inbred-midparent $F_2$ heterosis ($IMF_2H$) (Table 2.3). When populations start to be selfed to produce the inbred parent populations, IMPV starts to decrease as a consequence of the expression of several recessive alleles, however, once the population approaches its limit ($F_{ST} = 1$), IMPV will approach midparent heterosis. In the case of crosses involving only inbred lines, the value for midparent heterosis and inbred-midparent heterosis will be the same. An inbred line constitutes the limit that an individual can reach when inbred to homozygosity and therefore, the mean when the inbred parent is selfed ($f = \frac{1}{2}$) will be the same as the mean before selfing ($f = 0$).
**Average inbreeding depression in the parents (ID)**

Inbreeding depression refers to the reduction in the phenotypic mean because of inbreeding (Falconer and Mackay 1996). It is measured as the difference in performance between an outbred and an inbred population (Hedrick 2005). Lamkey and Edwards (1999) provide several and equivalent estimators of inbreeding depression. For them, inbreeding depression could be estimated as twice the difference between the reference population ($F_2$) and the selfed $F_1$ generation ($2F_2 - 2F_1$-selfed), as the deviation of the $F_2$ from IMPV (i.e. IMF$_2$H), or as the difference between inbred-midparent and midparent heterosis (IMH - MH). In the present study, we use the latter because of its simplicity. By substituting each term from table 2.3:

\[
ID = IMH - MH = (F_1 - IMPV) - (F_1 - MPV) = MPV - IMPV = 2\bar{p}_1\bar{p}_2d - 2\Delta^2d
\]

Where $2\bar{p}_1\bar{p}_2d$ refers to the change of mean caused by inbreeding and $2\Delta^2d$ to a function of dominance and the divergence in allele frequencies between parents ($\Delta$). Both effects can be further expressed in terms of the allele frequencies of each parent. Earlier, the average allele frequency was defined as $\bar{p}_1 = \frac{1}{2}(p_1 + p_1')$ and $\bar{p}_2 = \frac{1}{2}(p_2 + p_2')$, while the divergence in allele frequency was defined as $\Delta = \delta_1 = -\delta_2$. After expanding $2\bar{p}_1\bar{p}_2d$, it can be expressed as a function of the sum of $\frac{1}{4}(2p_1p_2d + 2p_1'p_2'd)$, where the terms within brackets refer to the sum of the inbreeding depression in the panmictic population 1 and 2, and $\frac{1}{4}(2p_1'p_2'd + 2p_1p_2d)$, where the terms within brackets refer to the dominance produced by the cross-product of allele frequencies. Likewise, $2\Delta^2d$ could be expanded to a linear function of one fourth the dominance produced by the cross-product minus one fourth the inbreeding depression suffered for each panmictic population ($\frac{1}{4}(2p_1'p_2d + 2p_1p_2'd) - \frac{1}{4}(2p_1p_2d + 2p_1'p_2'd)$). Therefore, the average inbreeding depression in the parents can be expressed as:

\[
ID = 2\bar{p}_1\bar{p}_2d - 2\Delta^2d = \frac{1}{2}(2p_1p_2d + 2p_1'p_2'd)
\]
**Results and discussion**

Based on Bonferroni adjusted p-values (data not shown), twenty-two, thirteen, and fourteen observations were identified as outliers and removed from the analysis for grain yield, plant height, and ear height, respectively. Bayesian information criterion (BIC) is given for each of the 24 fitted models in table 2.4. The best-fitting model allows heterogeneous residual variances among environments \( V(\epsilon_{ijklm}) = \sigma^2_{\epsilon(i)} \) for all traits (Table 2.4). In terms of heterogeneous range within replication and environment interaction \( \delta_{ni(l)} \), all traits except ear height allow heterogeneity among environments \( V(\delta_{ni(l)}) = \sigma^2_{\delta(i)} \) (Table 2.4). For grain yield, a model including all random effects was selected, while for the other two traits a model excluding the environment by inbreeding level interaction effect \( \omega_{ij} \) was best fitted to the data (Table 2.4).

**Heterosis and average inbreeding depression in parents**

Midparent heterosis (MH) and inbred-midparent heterosis (IMH), calculated from the genetic parameters of the model (Table 2.3), were significantly different from zero for all traits and crosses (Table 2.5). For inbred line crosses, estimates of midparent and inbred-midparent heterosis were statistically the same because an inbred line constitutes the limit that an individual can reach when inbred to homozygosis. For crosses involving a panmictic population, inbred-midparent heterosis was significantly higher than MH for all crosses and traits (Table 2.5). Our model predicts that midparent heterosis is a function of dominance and genetic divergence in allele frequencies between the parents \( \Delta \), while inbred-midparent heterosis depends on \( \Delta \), dominance, and inbreeding depression. Marsan et al. (1998), found a positive correlation between the genetic distance of the parents and hybrid performance, and Reif et al. (2003), concludes that when the genetic divergence in allele frequency between the parent increases, an increase in the probability of complementation of favorable genes produces heterosis. During the
process of inbreeding of a panmictic population, the number of effective parents ($N_e$) is reduced, resulting in an increased probability of crossing individuals related by ancestry (Falconer and Mackay 1996). As the panmictic populations become inbred (increased $F_{ST}$), a decrease in the levels of heterozygosity within populations, and an increase in the genetic divergence between populations is obtained. In that regard, Edwards and Lamkey (2003) concluded that by increasing the level of inbreeding, a significant reduction in the genetic diversity within populations is obtained (because of low heterozygosity), while the variance among populations is increased almost linearly when heterozygosity within populations is reduced. The decrease in heterozygosity within a population results in the homozygosis of different recessive deleterious alleles that will cause a decline in fitness, reducing the parent population means (Charlesworth and Charlesworth 1999). In general, crosses involving inbred lines (e.g. population-by-inbred and inbred-by-inbred) showed lower midparent values and higher midparent heterosis because of a reduced level of heterozygosity in the parents (Figure 2.2). Therefore, higher midparent heterosis is related to lower midparent values, both influenced by the heterozygosity within populations. For instance, for grain yield, there was a decrease of $1.42 \pm 0.17 \text{ Mg ha}^{-1}$ in midparent heterosis per one Mg increase in midparent values, while for plant and ear height there was a decrease of $0.96 \pm 0.35$ and $0.75 \pm 0.36 \text{ cm}$ per each centimeter of increase in midparent values (Figure 2.2).

Inbreeding depression has been defined as being dependent on the level of dominance and genotype frequencies (Falconer and Mackay 1996; Lynch and Walsh 1998). During inbreeding, the frequency of homozygotes is increased at the expense of the heterozygotes, and thus, a change in the population mean after inbreeding could be associated with reduced heterozygosity in the population (Falconer and Mackay 1996; Lynch and Walsh 1998).
Therefore, assuming there is dominance, lower inbreeding depression will be obtained as the population approaches complete inbreeding ($F_{ST} = 1$) because of a reduction in heterozygosity. The limiting case will be obtained when the population becomes an inbred line, in which case there can be no further change in the population mean with further inbreeding. Increasing $F_{ST}$ results in increasing heterosis because of increasing genetic divergence between parents ($\Delta$) and low inbreeding depression within parent populations because of reduced heterozygosity. Consequently, low values for inbreeding depression in the parents (obtained at high $F_{ST}$), are related to high midparent heterosis (Figure 2.3). Our results showed that for grain yield, there was a decrease of $0.42 \pm 0.12$ Mg ha$^{-1}$ in midparent heterosis for each one Mg increase in inbreeding depression in the parent populations, while for plant and ear height, there was a decrease of $0.74 \pm 0.26$ cm and $0.79 \pm 0.09$ cm respectively per each centimeter of increase in inbreeding depression in the parent populations (Figure 2.3). For all traits, crosses involving an inbred line (population-by-inbred) showed the lowest values for the average inbreeding depression in the parents and the highest values for midparent heterosis (Figure 2.3). Hence, there is a tradeoff between inbreeding depression (or midparent values) and heterosis, revealing that by increasing the genetic divergence in allele frequency between parents ($\Delta$) the inbreeding depression, midparent value, and heterozygosity is reduced.

Genetic importance of a hybrid seed production system

During the past 50 to 70 years, the agronomic performance of maize has been primarily improved using $F_1$ hybrids. However, a loss in performance when saving seed from $F_1$ hybrids have been reported (Hallauer and Miranda 1988; Meghji et al. 1984). This loss could be attributed to a reduction of half heterosis observed in the $F_1$ or $2\Delta^2d$ (Falconer and Mackay 1996; Lamkey and Edwards 1999). Our results showed an increased performance when using an $F_1$
compared to F₂ hybrids (Table 2.7). Crosses involving only inbred lines showed a higher loss, while the lower loss was obtained in the population-by-population crosses (Table 2.7). According to Hallauer and Miranda (1988), reductions in the mean of the F₂ are smaller when open-pollinated varieties are used as parents compared to cases where inbred lines are used. For open-pollinated varieties with unknown allele frequencies, estimates for Δ are expected to be less than half (½), and thus the loss in performance from saving F₁ hybrids seed (2Δ²d) is expected to be less than ½d as is the case for single-inbred line crosses (Table 2.3). Equally, Lynch (1991) mentioned that the loss in fitness between the F₁ and F₂ is greater when the two parents are genetically distant, as is the case for inbred-by-inbred crosses in our experiment. Saving seed from an F₁ hybrid corresponds to a reduction in performance attributable to half inbred-midparent heterosis or ½(2p₁p₂d + 2Δ²d). Therefore, the value of a hybrid seed production system is, as pointed by Shull (1909), to find and maintain the best genotype, which will express an increment of half midparent heterosis (or half inbred midparent heterosis) when compared to an open-pollinated variety.

Summary

For the twelve different crosses used in the present experiment, a single-locus model with two alleles was reasonably good to explain the mean of each generation and to predict heterosis. According to the model, heterosis could be explained in terms of dominance (d), genetic divergence in allele frequencies (Δ), and inbreeding depression. Midparent heterosis (MH) could be explained by the dominance and the genetic divergence in allele frequency between the parents (4Δ²d), while inbred-midparent heterosis depends on dominance, Δ, and inbreeding depression (Table 2.3). Both empirical and model-based evidence show that the genetic divergence in allele frequency between parents is the key driver of heterosis, with inbred-by-
inbred crosses being the limiting case. Genetic differentiation in allele frequencies between parent populations is achieved when the level of heterozygosity within each population decrease (increased F\textsubscript{ST}) and several deleterious alleles are fixed in each parent. Similarly, at high levels of genetic differentiation between parents, there will be a low effect of inbreeding depression in the populations because a significant reduction of heterozygosity. Therefore, heterosis could be explained as the result of inbreeding the parents, which will result in a high genetic divergence between parents and low inbreeding depression. Finally, losses in agronomic performance caused by saving seed from the field were calculated to be one-half midparent heterosis. In that regard, the genetic value of a hybrid seed production system over interpopulation crosses or offspring from non-inbred parents is attributable to a gain in half midparent heterosis.

**Tables and Figures**

Table 2.1. Derivation process and development of the fourteen (14) parents evaluated in the experiment.

<table>
<thead>
<tr>
<th>Pedigree</th>
<th>Derivation</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>BSSS(R)C15</td>
<td>The fifteenth cycle of reciprocal recurrent selection with BSCB1 as a tester in the selection scheme</td>
<td>Penny and Eberhart (1971)</td>
</tr>
<tr>
<td>BSCB1(R)C15</td>
<td>The fifteenth cycle of reciprocal recurrent selection with BSSS as a tester in the selection scheme</td>
<td>Penny and Eberhart (1971)</td>
</tr>
<tr>
<td>BSCB1(R)C16</td>
<td>The sixteenth cycle of reciprocal recurrent selection with BSSS as a tester in the selection scheme</td>
<td>Penny and Eberhart (1971)</td>
</tr>
<tr>
<td>BS13(S)C10</td>
<td>The tenth cycle of an S\textsubscript{2}-progeny selection from the recombination of 29 S\textsubscript{1} lines</td>
<td>Lamkey (1992)</td>
</tr>
<tr>
<td>BS11(FR)C15</td>
<td>The fifteenth cycle of a reciprocal full-sib recurrent selection (FR) between BS11 and BS10</td>
<td>Weyhrich et al. (1998)</td>
</tr>
<tr>
<td>BSKRL4(HI)C2</td>
<td>The second cycle of a half-sib recurrent selection program started at Iowa State University</td>
<td>Personal communication Dr. Lamkey 2016</td>
</tr>
<tr>
<td>B114</td>
<td>Line developed based on testcross performance from “Pool 41-C15-19-2-1-1-1-1-1-1” with the tester A632</td>
<td>Hallauer et al. (2000)</td>
</tr>
<tr>
<td>B116</td>
<td>Line from the sixth generation of pedigree selection applied to the B97/B99 F\textsubscript{2} population</td>
<td>Hallauer et al. (2004)</td>
</tr>
<tr>
<td>B129</td>
<td>Line produce by single seed descent (SSD) from the cross B73/B84</td>
<td></td>
</tr>
<tr>
<td>B73</td>
<td>Line selected in the fifth cycle of the reciprocal recurrent selection with BSCB1 as a common tester</td>
<td>Russell (1972)</td>
</tr>
<tr>
<td>BX010</td>
<td>An S\textsubscript{12} experimental line developed from the cross B95/B97</td>
<td>Hallauer et al. (1992)</td>
</tr>
<tr>
<td>Mo17</td>
<td>Line selected from the single cross CI187-2/C103 in the Missouri Agricultural Experiment Station</td>
<td>Zuber (1973)</td>
</tr>
</tbody>
</table>
Table 2.1 Continued

<table>
<thead>
<tr>
<th>Pedigree</th>
<th>Derivation</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>SGI912</td>
<td>Line from by Seed Genetics Inc. from the cross B73/B37</td>
<td></td>
</tr>
<tr>
<td>TR7245</td>
<td>Line developed by Thurston Genetics</td>
<td></td>
</tr>
</tbody>
</table>

Table 2.2. Crosses made between the synthetic populations, between synthetic populations and the inbred line B129, and between inbred lines.

<table>
<thead>
<tr>
<th>Cross</th>
<th>Parent 1</th>
<th>Parent 2</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>BS13(S)C10</td>
<td>BSCB1(R)C15</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>BS13(S)C10</td>
<td>BSSS(R)C15</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>BSSS(R)C15</td>
<td>BSCB1(R)C15</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>BS11(FR)C15</td>
<td>B129</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>BSCB1(R)C16</td>
<td>B129</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>BSKRL4(HI)C2</td>
<td>B129</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>B129</td>
<td>B114</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>B129</td>
<td>B116</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>B73</td>
<td>Mo17</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>SGI912</td>
<td>BX010</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>SGI912</td>
<td>B116</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>TR7245</td>
<td>B116</td>
<td></td>
</tr>
</tbody>
</table>

Table 2.3. Genetic expected means for each generation, midparent value at panmixia (MPV) and complete inbreeding (IMPV), and heterosis for panmictic populations and for inbred-by-inbred crosses.

<table>
<thead>
<tr>
<th>Generation/Estimate</th>
<th>Population crosses</th>
<th>Inbred by Inbred</th>
</tr>
</thead>
<tbody>
<tr>
<td>F2</td>
<td>µ</td>
<td>½d</td>
</tr>
<tr>
<td>F1</td>
<td>µ + 2Δ^2d</td>
<td>d</td>
</tr>
<tr>
<td>F1-selfed</td>
<td>µ − p1p2d + Δ^2d</td>
<td>½d</td>
</tr>
<tr>
<td>P1 (µ1)</td>
<td>µ − 2Δ^2d + 2α</td>
<td>a</td>
</tr>
<tr>
<td>P2 (µ1)</td>
<td>µ − 2Δ^2d − 2α</td>
<td>−a</td>
</tr>
<tr>
<td>P1 (µ1) / Inbred1</td>
<td>µ − 2(p1p2d + 2Δa)</td>
<td>a</td>
</tr>
<tr>
<td>P2 (µ1) / Inbred2</td>
<td>µ − 2(p1p2d − 2Δa)</td>
<td>−a</td>
</tr>
<tr>
<td>BC_P1/BC_Inbred1</td>
<td>µ + Δα</td>
<td>½d + ½a</td>
</tr>
<tr>
<td>BC_P2/BC_Inbred2</td>
<td>µ − Δα</td>
<td>½d − ½a</td>
</tr>
<tr>
<td>MPV</td>
<td>µ − 2Δ^2d</td>
<td>0</td>
</tr>
<tr>
<td>MH</td>
<td>4Δ^2d</td>
<td>d</td>
</tr>
<tr>
<td>IMH</td>
<td>2p1p2d + 2Δ^2d</td>
<td>d</td>
</tr>
<tr>
<td>MF2_H</td>
<td>2Δ^2d</td>
<td>½d</td>
</tr>
<tr>
<td>IMF2_H</td>
<td>2p1p2d</td>
<td>½d</td>
</tr>
<tr>
<td>ID</td>
<td>2(p1p2d − 2Δ^2) = ½(2p1p2d + 2p1’p2’d)</td>
<td>0</td>
</tr>
</tbody>
</table>

MH: midparent heterosis; IMH: inbred-midparent heterosis; MF2_H: midparent F2 heterosis; IMF2_H: inbred-midparent F2 heterosis; MPV: midparent value; IMPV: inbred-midparent value; α: additive effect at a locus; d: dominant effect at a locus; p1 and p2: average allele frequency for allele 1 and 2 respectively; a: average effect of an allele substitution (α + (p2 − p1)d); µ: population mean (α(p1 − p2) + 2p1*p2’d).
Table 2.4. Bayesian information criteria (BIC) for the 24 possible linear models. The models provided contained all fixed effects described in the materials and methods section plus the random effects containing an ‘X’ in the corresponding column. The lowest BIC value is denoted with an asterisk.

<table>
<thead>
<tr>
<th>Env x Inb</th>
<th>Env x Ped(Inb)</th>
<th>Range (Rep x Env)</th>
<th>Het Error†</th>
<th>Het Range‡</th>
<th>Grain Yield</th>
<th>Plant Height</th>
<th>Ear Height</th>
</tr>
</thead>
<tbody>
<tr>
<td>X</td>
<td>X</td>
<td>X</td>
<td>N</td>
<td>Y</td>
<td>3298.6</td>
<td>7541.7</td>
<td>6736.1</td>
</tr>
<tr>
<td>X</td>
<td>X</td>
<td>X</td>
<td>N</td>
<td>N</td>
<td>3342.4</td>
<td>7563.2</td>
<td>6716.2</td>
</tr>
<tr>
<td>X</td>
<td>X</td>
<td>X</td>
<td>Y</td>
<td>N</td>
<td>3186.6</td>
<td>7482.8</td>
<td>6702.3</td>
</tr>
<tr>
<td>X</td>
<td>X</td>
<td>X</td>
<td>Y</td>
<td>Y</td>
<td>3181.2</td>
<td>7460.8</td>
<td>6729.3</td>
</tr>
<tr>
<td>X</td>
<td>X</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>3462.3</td>
<td>7675.4</td>
<td>6736.0</td>
</tr>
<tr>
<td>X</td>
<td>X</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>3245.6</td>
<td>7594.6</td>
<td>6721.9</td>
</tr>
<tr>
<td>X</td>
<td>X</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td>3434.7</td>
<td>7595.1</td>
<td>6764.6</td>
</tr>
<tr>
<td>X</td>
<td>X</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>3458.0</td>
<td>7614.5</td>
<td>6746.1</td>
</tr>
<tr>
<td>X</td>
<td>X</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>3285.1</td>
<td>7556.8</td>
<td>6726.2</td>
</tr>
<tr>
<td>X</td>
<td>X</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>3280.7</td>
<td>7532.9</td>
<td>6752.3</td>
</tr>
<tr>
<td>X</td>
<td>X</td>
<td>N</td>
<td>Y</td>
<td>N</td>
<td>3308.1</td>
<td>7534.5</td>
<td>6729.2</td>
</tr>
<tr>
<td>X</td>
<td>X</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>3352.9</td>
<td>7556.0</td>
<td>6709.3</td>
</tr>
<tr>
<td>X</td>
<td>X</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>3200.2</td>
<td>7475.6</td>
<td>6695.6</td>
</tr>
<tr>
<td>X</td>
<td>X</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>3194.0</td>
<td>7453.7</td>
<td>6722.6</td>
</tr>
<tr>
<td>X</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>3547.1</td>
<td>7718.3</td>
<td>6762.0</td>
</tr>
<tr>
<td>X</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>3338.3</td>
<td>7661.0</td>
<td>6742.2</td>
</tr>
<tr>
<td>X</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>3495.1</td>
<td>7668.8</td>
<td>6730.0</td>
</tr>
<tr>
<td>X</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>3273.0</td>
<td>7587.4</td>
<td>6716.4</td>
</tr>
<tr>
<td>X</td>
<td>N</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>3462.1</td>
<td>7587.9</td>
<td>6759.1</td>
</tr>
<tr>
<td>X</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>3488.5</td>
<td>7607.3</td>
<td>6740.0</td>
</tr>
<tr>
<td>X</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>3315.5</td>
<td>7549.7</td>
<td>6721.0</td>
</tr>
<tr>
<td>X</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>3311.3</td>
<td>7525.8</td>
<td>6746.8</td>
</tr>
<tr>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>3649.7</td>
<td>7717.2</td>
<td>6759.4</td>
</tr>
<tr>
<td>Y</td>
<td>N</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>3415.7</td>
<td>7660.3</td>
<td>6742.0</td>
</tr>
</tbody>
</table>

Env: environment; Rep: replication; Ped: pedigree; Inb: inbreeding level; † indicates if the model contained heterogeneous residual variance among environments ($V(\epsilon_{ijktn}) = \sigma^2_\varepsilon$ or $V(\epsilon_{ijktn}) = \sigma^2_{\varepsilon(i)}$); ‡ indicates if the model contained heterogeneous range within replication by environment interaction variance among environments ($V(\delta_{n(i)}) = \sigma^2_\delta$ or $V(\delta_{n(i)}) = \sigma^2_{\delta(i)}$); N: a model with homogeneous variance for the effect; Y: a model with heterogeneous variance for the effect.
Table 2.5. Mean values (standard error) for midparent value (MPV), inbred midparent value (IMPV), midparent heterosis (MH), inbred-midparent heterosis (IMH), midparent $F_2$ ($MF_2H$) heterosis, and inbred-midparent $F_2$ ($IMF_2H$) heterosis for ear height (cm), plant height (cm), and grain yield (Mg ha$^{-1}$) for the inbred-by-inbred (IxI), population-by-inbred (PxI), and population-by-population (PxP) crosses.

<table>
<thead>
<tr>
<th>Pedigree</th>
<th>Cross</th>
<th>MPV</th>
<th>IMPV</th>
<th>MH</th>
<th>IMH</th>
<th>$MF_2H$</th>
<th>$IMF_2H$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Grain Yield (Mg ha$^{-1}$)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B129/B114</td>
<td>IxI</td>
<td>3.73 ± 0.48</td>
<td>3.41 ± 0.23</td>
<td>6.76 ± 0.68</td>
<td>7.08 ± 0.39</td>
<td>3.38 ± 0.34</td>
<td>3.70 ± 0.30</td>
</tr>
<tr>
<td>B129/B116</td>
<td>IxI</td>
<td>5.46 ± 0.68</td>
<td>3.94 ± 0.27</td>
<td>4.81 ± 1.00</td>
<td>6.33 ± 0.53</td>
<td>2.40 ± 0.50</td>
<td>3.92 ± 0.38</td>
</tr>
<tr>
<td>B73/Mo17</td>
<td>IxI</td>
<td>3.89 ± 0.86</td>
<td>4.44 ± 0.42</td>
<td>5.13 ± 1.21</td>
<td>4.57 ± 0.70</td>
<td>2.56 ± 0.61</td>
<td>2.01 ± 0.55</td>
</tr>
<tr>
<td>SGI912/B116</td>
<td>IxI</td>
<td>3.99 ± 0.42</td>
<td>4.78 ± 0.18</td>
<td>6.98 ± 0.60</td>
<td>6.19 ± 0.32</td>
<td>3.49 ± 0.30</td>
<td>2.70 ± 0.25</td>
</tr>
<tr>
<td>SGI912/BX010</td>
<td>IxI</td>
<td>3.85 ± 0.96</td>
<td>4.79 ± 0.45</td>
<td>6.80 ± 1.41</td>
<td>5.86 ± 0.80</td>
<td>3.40 ± 0.70</td>
<td>2.46 ± 0.58</td>
</tr>
<tr>
<td>TR7245/B116</td>
<td>IxI</td>
<td>4.96 ± 0.56</td>
<td>4.82 ± 0.26</td>
<td>5.47 ± 0.79</td>
<td>5.61 ± 0.44</td>
<td>2.73 ± 0.40</td>
<td>2.88 ± 0.34</td>
</tr>
<tr>
<td>BS11(FR)C15/B129</td>
<td>PxI</td>
<td>5.75 ± 0.46</td>
<td>3.23 ± 0.43</td>
<td>3.99 ± 0.70</td>
<td>6.51 ± 0.59</td>
<td>1.99 ± 0.35</td>
<td>4.51 ± 0.50</td>
</tr>
<tr>
<td>BSCB1(R)C16/B129</td>
<td>PxI</td>
<td>5.36 ± 0.18</td>
<td>3.24 ± 0.17</td>
<td>4.28 ± 0.28</td>
<td>6.41 ± 0.23</td>
<td>2.14 ± 0.14</td>
<td>4.27 ± 0.20</td>
</tr>
<tr>
<td>BSKR14(HJ)C2/B129</td>
<td>PxI</td>
<td>4.88 ± 0.24</td>
<td>3.22 ± 0.23</td>
<td>4.88 ± 0.36</td>
<td>6.54 ± 0.31</td>
<td>2.44 ± 0.18</td>
<td>4.10 ± 0.27</td>
</tr>
<tr>
<td>BS13(S)C10/BS13(S)C10</td>
<td>PxP</td>
<td>6.70 ± 0.26</td>
<td>4.17 ± 0.54</td>
<td>2.31 ± 0.48</td>
<td>4.85 ± 0.67</td>
<td>1.15 ± 0.24</td>
<td>3.69 ± 0.62</td>
</tr>
<tr>
<td>BS13(S)C10/BSB1(R)C15</td>
<td>PxP</td>
<td>7.24 ± 0.10</td>
<td>4.63 ± 0.20</td>
<td>1.60 ± 0.18</td>
<td>4.21 ± 0.25</td>
<td>0.80 ± 0.09</td>
<td>3.41 ± 0.23</td>
</tr>
<tr>
<td>BS13(S)C10/BS13(S)C10</td>
<td>PxP</td>
<td>6.12 ± 0.27</td>
<td>2.92 ± 0.57</td>
<td>2.58 ± 0.50</td>
<td>5.78 ± 0.70</td>
<td>1.29 ± 0.25</td>
<td>4.49 ± 0.65</td>
</tr>
<tr>
<td>BS11(FR)C15/B129</td>
<td>PxP</td>
<td>104.71 ± 3.46</td>
<td>88.22 ± 3.15</td>
<td>7.81 ± 5.25</td>
<td>24.30 ± 4.32</td>
<td>3.91 ± 2.62</td>
<td>20.40 ± 3.73</td>
</tr>
<tr>
<td>BSCB1(R)C16/B129</td>
<td>PxP</td>
<td>88.28 ± 1.81</td>
<td>75.62 ± 1.65</td>
<td>16.33 ± 2.75</td>
<td>28.79 ± 2.27</td>
<td>8.17 ± 1.38</td>
<td>20.62 ± 1.95</td>
</tr>
<tr>
<td>BS13(S)C10/BSB1(R)C15</td>
<td>PxP</td>
<td>93.75 ± 1.60</td>
<td>82.24 ± 1.46</td>
<td>18.38 ± 2.44</td>
<td>29.89 ± 2.04</td>
<td>9.19 ± 1.22</td>
<td>20.70 ± 1.72</td>
</tr>
<tr>
<td>BS13(S)C10/BS13(S)C10</td>
<td>PxP</td>
<td>84.65 ± 0.89</td>
<td>74.54 ± 1.85</td>
<td>16.05 ± 1.70</td>
<td>26.16 ± 2.32</td>
<td>8.03 ± 0.85</td>
<td>18.13 ± 2.15</td>
</tr>
<tr>
<td>BS13(S)C10/BS13(S)C10</td>
<td>PxP</td>
<td>93.26 ± 1.05</td>
<td>79.10 ± 2.17</td>
<td>11.60 ± 1.98</td>
<td>25.76 ± 2.74</td>
<td>5.80 ± 0.99</td>
<td>19.96 ± 2.53</td>
</tr>
<tr>
<td>BS13(S)C10/BS13(S)C10</td>
<td>PxP</td>
<td>91.09 ± 4.57</td>
<td>76.04 ± 9.41</td>
<td>14.46 ± 8.62</td>
<td>29.52 ± 11.88</td>
<td>7.23 ± 4.31</td>
<td>22.29 ± 10.99</td>
</tr>
<tr>
<td>Pedigree</td>
<td>Cross</td>
<td>MPV</td>
<td>IMPV</td>
<td>MH</td>
<td>IMH</td>
<td>MF$_2$H</td>
<td>IMF$_2$H</td>
</tr>
<tr>
<td>---------------</td>
<td>-------</td>
<td>---------</td>
<td>----------</td>
<td>------</td>
<td>---------</td>
<td>---------</td>
<td>---------</td>
</tr>
<tr>
<td>B129/B114</td>
<td>lxl</td>
<td>178.46 ± 5.67</td>
<td>170.53 ± 2.30</td>
<td>44.96 ± 7.88</td>
<td>52.88 ± 4.13</td>
<td>22.48 ± 3.94</td>
<td>30.40 ± 3.25</td>
</tr>
<tr>
<td>B129/B116</td>
<td>lxl</td>
<td>200.30 ± 13.54</td>
<td>194.47 ± 4.95</td>
<td>34.23 ± 19.76</td>
<td>40.07 ± 10.17</td>
<td>17.12 ± 9.88</td>
<td>22.95 ± 7.29</td>
</tr>
<tr>
<td>B73/Mo17</td>
<td>lxl</td>
<td>187.70 ± 2.69</td>
<td>185.73 ± 1.28</td>
<td>47.15 ± 3.75</td>
<td>49.12 ± 2.14</td>
<td>23.58 ± 1.88</td>
<td>25.54 ± 1.70</td>
</tr>
<tr>
<td>SGi912/B116</td>
<td>lxl</td>
<td>194.90 ± 3.14</td>
<td>191.83 ± 1.23</td>
<td>54.46 ± 4.48</td>
<td>57.53 ± 2.34</td>
<td>27.23 ± 2.24</td>
<td>30.30 ± 1.77</td>
</tr>
<tr>
<td>SGi912/BX010</td>
<td>lxl</td>
<td>189.25 ± 2.22</td>
<td>184.32 ± 0.92</td>
<td>58.01 ± 3.29</td>
<td>62.95 ± 1.79</td>
<td>29.01 ± 1.64</td>
<td>33.94 ± 1.26</td>
</tr>
<tr>
<td>TR7245/B116</td>
<td>lxl</td>
<td>199.03 ± 4.28</td>
<td>196.42 ± 1.94</td>
<td>44.03 ± 6.06</td>
<td>46.64 ± 3.39</td>
<td>22.02 ± 3.03</td>
<td>24.63 ± 2.63</td>
</tr>
<tr>
<td>BS11(FR)C15/B129</td>
<td>Pxl</td>
<td>211.79 ± 4.41</td>
<td>184.10 ± 4.04</td>
<td>15.83 ± 6.71</td>
<td>43.52 ± 5.58</td>
<td>7.91 ± 3.36</td>
<td>35.60 ± 4.78</td>
</tr>
<tr>
<td>BSCB1(R)C16/B129</td>
<td>Pxl</td>
<td>191.54 ± 6.56</td>
<td>171.28 ± 6.04</td>
<td>33.22 ± 9.99</td>
<td>53.47 ± 8.24</td>
<td>16.61 ± 5.00</td>
<td>36.87 ± 7.12</td>
</tr>
<tr>
<td>BSKRL4(HI)C2/B129</td>
<td>Pxl</td>
<td>201.96 ± 3.76</td>
<td>183.45 ± 3.45</td>
<td>30.64 ± 5.74</td>
<td>49.15 ± 4.79</td>
<td>15.32 ± 2.87</td>
<td>33.83 ± 4.05</td>
</tr>
<tr>
<td>BS13(S)C10/BSCB1(R)C15</td>
<td>PxP</td>
<td>187.94 ± 2.06</td>
<td>166.40 ± 4.20</td>
<td>36.35 ± 3.81</td>
<td>57.89 ± 5.28</td>
<td>18.18 ± 1.91</td>
<td>39.72 ± 4.90</td>
</tr>
<tr>
<td>BS13(S)C10/BSSS(R)C15</td>
<td>PxP</td>
<td>200.02 ± 1.24</td>
<td>175.61 ± 2.52</td>
<td>20.56 ± 2.29</td>
<td>44.96 ± 3.17</td>
<td>10.28 ± 1.14</td>
<td>34.68 ± 2.94</td>
</tr>
<tr>
<td>BSSS(R)C15/BSCB1(R)C15</td>
<td>PxP</td>
<td>204.55 ± 6.22</td>
<td>176.04 ± 12.62</td>
<td>26.04 ± 11.51</td>
<td>54.54 ± 15.85</td>
<td>13.02 ± 5.76</td>
<td>41.52 ± 14.73</td>
</tr>
</tbody>
</table>
Figure 2.1. Mating scheme for the panmictic populations ($P_1$ and $P_2$) with average inbreeding compared to a metapopulation bigger than zero ($F_{ST} > 0$). Panmictic populations are inbred for several generations to create a population of inbreds. Genetically, crossing the two populations of inbreds will produce the same $F_1$ as crossing the two panmictic populations. Random mating individuals in the $F_1$ gives origin to the $F_2$, considered as the reference population.
Figure 2.2. Simple linear regression of midparent heterosis versus midparent values for all crosses and A) grain yield, B) plant height, and C) ear height. Dots correspond to inbred-by-inbred crosses, upward triangle to population-by-inbred crosses, and the downward triangles to population-by-population crosses.
Figure 2.3. Simple linear regression of inbreeding depression in the parent populations versus midparent heterosis for A) grain yield, B) plant height, and C) ear height. Upward triangles represent the population-by-inbred cross, while downward triangles represent the population-by-population crosses.

References


Jones DF (1917) Dominance of linked factors as a means of accounting for heterosis. Genetics 2:466-479


Kempthorne O (1957) An introduction to genetic statistics


Shull GH (1914) Duplicate genes for capsule-form in Bursa bursa-pastoris. Z Ind Abst Ver 12:97-149


CHAPTER 3. GENETIC COMPOSITION AND GENETIC DIVERSITY OF 44 EXPIRED PLANT VARIETY PROTECTION (ex-PVP) MAIZE INBRED LINES THROUGH HIGH-RESOLUTION DETECTION OF IDENTITY BY DESCENT SEGMENTS

Fernando S. Aguilar\textsuperscript{1}, Mark A. Mikel\textsuperscript{2}, and Jode W. Edwards\textsuperscript{3}

\textsuperscript{1} Department of Agronomy, Iowa State University, Ames, IA 50011, USA

\textsuperscript{2} Department of Crop Sciences, University of Illinois, Carl R. Woese Institute for Genomic Biology

\textsuperscript{3} USDA-ARS Corn Insects and Crop Genetics Research Unit, Ames, IA 50011, USA

Modified from a manuscript to be submitted to Theoretical and Applied Genetics

Abstract

An IBD segment refers to segments of DNA descended from common ancestors. The identification of IBD segments could be used for the estimation of genetic diversity and genetic contribution. The main purpose of this study was i) to quantify the size of IBD segments, ii) to estimate the number of generations from a common ancestor for each segment, iii) to quantify the genetic diversity of 44 ex-PVP and eight key ancestor lines when computed from IBD segments, iv) to compare diversity estimates calculated from pedigree information, IBS probabilities, and IBD segments, and v) to quantify the genetic contribution of the eight key ancestors. Kinship coefficients were calculated from detected IBD segments using an HMM framework and from pedigree information. Long IBD segments (>14.8 Mb) were predominant between ex-PVP and key ancestor lines in both stiff stalk and non-stiff stalk heterotic groups. The length of IBD segments suggests an inbreeding event created no more than 15 generations ago. Kinship coefficients computed from IBD probabilities allows the estimation of relationships even under incomplete pedigree information, increasing the ability to find hidden relationships. There was a higher genetic contribution of the key ancestors B14, B37, and B73 to the lines of
the SSS group, compared to the contribution of Mo17, Oh43, and PH207 in the NSS group. We found that a small group of ancestors has contributed large proportions of the genome to important PVP lines, with lines in the SSS group having longer regions in the genome inherited from B73 and B14.

**Introduction**

Inbreeding has been used in quantitative genetics to denote the mating between individuals with shared ancestors (Bernardo 2010; Falconer and Mackay 1996; Malécot 1948; Rudan and Campbell 2004). Consequences of inbreeding are an increase in the probability that any two alleles at a locus are identical by descent (IBD) (Bernardo 2010; Falconer and Mackay 1996; Rudan and Campbell 2004). IBD occurs when identical alleles are inherited from a common ancestor (Hartl et al. 1997; Wright 1922) and constitutes a measure of the degree of relationship between and/or within individuals. Wright (1921), defines this relationship as the correlation between uniting gametes, while Malécot (1948) defines it as the probability that 2 homologous genes at a locus are IBD. In both definitions, the estimation of the degree of relationship (i.e. kinship coefficient) depends on the description of an ancestral population, which by definition is assumed to be the base from where past ancestry is no longer accounted (Falconer and Mackay 1996; Wright 1922). Thus, the lower the number of common ancestors or the lower the number of generations separating the ancestral with the current population, the higher the kinship coefficient between individuals because of a reduced number of possible recombination events (Wang 2014). During meiosis, DNA segments in the genome are broken as the result of the recombination between homologous chromosomes, resulting in smaller segments shared by descent (Fisher 1954; Stam 1980; Thompson 2013). Browning and Browning (2012) and Thompson (2013), suggest that large IBD segments are inherited from recent ancestors, while smaller segments are related to ancient inbreeding because of an elevated
number of recombination events. According to Thompson (2013), the non-random association of alleles at two or more loci (i.e. linkage disequilibrium, LD), is another factor influencing the inheritance of large or small segments of DNA and therefore can be viewed as a sign of coancestry between individuals. Knowledge about the relationship between individuals has been used by breeders to reduce inbreeding within their breeding programs and thus, avoid the negative consequences of mating closely related individuals. For instance, kinship coefficients have been used to estimate the genetic diversity within breeding pools and to estimate the genetic contribution of a set of parents to its descendants. Therefore, the estimation of kinship coefficients represents a way to better utilize breeding resources and to identify the best combination of parents to maximize hybrid performance (Beckett et al. 2017; Zhang et al. 2018).

Commonly, IBD is estimated with the use of pedigree information, however, with the advent of high-throughput genotyping technologies, it can be estimated at a molecular scale by the estimation of IBD segments (Curik et al. 2014). An IBD segment refers to a segment of DNA descended from a common ancestor and in which there are associations between closely spaced loci because of coancestry. Browning and Browning (2012), state that IBD segments in the genome can be detected using genome-wide genotypic data through either a genotype counting algorithm or a probabilistic model. The first corresponds to an algorithm that looks for matches along each chromosome by moving a user-defined sliding window (Browning and Browning 2012; Ceballos et al. 2018). This type of methodology called an IBD segment based on the number of consecutive homozygote SNPs exceeding a user-defined threshold (Browning and Browning 2012; Ceballos et al. 2018). The second methodology corresponds to a model-based probabilistic approach using hidden Markov models (HMM) in the estimation of the IBD status (Browning and Browning 2012; Ceballos et al. 2018). HMM works as a Markov chain in which
the internal states (IBD status of each SNP) are estimated using a prior, transition, and emission probabilities. The prior probability describes the odds of IBD for the first marker, while the transition probability describes the likelihood of transitioning from the IBD status of maker $k - 1$ to maker $k$ (Han and Abney 2011; Leutenegger et al. 2003; Westhead and Vijayabaskar 2017). Finally, the emission probability describes the likelihood of the observed genotype at marker $k$ given the hidden state of that marker (Westhead and Vijayabaskar 2017). When high-density SNP data sets are used, additional care about linkage and linkage disequilibrium (LD) between markers must be considered. The incorporation of LD in the HMM framework has resulted in complex models where the emission probability becomes a conditional likelihood of the observed genotype at a maker given the genotype of adjacent loci (Browning and Browning 2012; Han and Abney 2011, 2013).

Identification of shared segments in the genome has been used for a range of purposes including the quantification of inbreeding, identification of patterns of inheritance, genotype imputation, haplotype phasing, genetic diversity analysis, and genetic contribution of a set of progenitors to its descendants (Albrechtsen et al. 2009; Browning and Browning 2007; Ferenčaković et al. 2013; Keller et al. 2011; Kirin et al. 2010; Lencz et al. 2007; Nelson et al. 2008a). Coffman et al. (2019) state that estimating the genetic contribution and diversity of ancestors to new germplasm could increase the understanding of public and private germplasm structure as well as provide information about pre-commercial and commercial maize lines. Several studies have reported the use of segments identical by state (IBS) in genome-wide association studies (GWAS), genomic prediction (Ferdosi et al. 2016; Schrag et al. 2007), identification of population structure, and to improve estimation on relationships between individuals (Gattepaille and Jakobsson 2012; Haasl and Payseur 2011; Lawson et al. 2012; Ralph
and Coop 2013), but few were used to identify the genetic contribution of ancestors to descendants in maize (Coffman et al. 2019; Jiao et al. 2012; Romero-Severson et al. 2001; Wu et al. 2016). Commonly, genotypic data have been used to estimate genetic diversity and ancestor’s genetic contribution using cluster analysis in the identification of shared identical by state segments. For instance, Coffman et al. (2019) divided the genome into haplotype groups by selecting a fixed SNP window that contains a minimum of 2000 SNPs, maximum diversity of 3%, maximum heterozygosity of 5%, and a minimum marker missing rate of 60%. Within each haplotype group, the length of shared IBS haplotype blocks was estimated between 174 ex-PVP and 60 ancestor lines and then divided by the total haplotype length for each ex-PVP (Coffman et al. 2019). In their analysis, Coffman et al. (2019) found a significant genetic contribution of the lines B73, B14, and B37 to inbred lines from a stiff stalk (SSS) heterotic group. Equally, they found that the line A632 ([(Mt42/1/B14)/2/B14/3/B14/4/B14]) had the fourth-highest amount of shared IBS haplotypes with the 174 ex-PVP lines and that the genetic contribution of B14 to this line was 78.7%. Nelson et al. (2008a), generated a Jaccard’s similarity coefficient matrix in which the proportion of shared IBS markers between 92 ex-PVP and 17 ancestor lines was used to create a dendrogram with an unweighted pair group method with arithmetic mean (UPGMA). The analysis grouped the 92 ex-PVP lines in six predominant clusters represented by the lines B73, Mo17, PH207, A632, Oh43, and B37 (Nelson et al. 2008a). Similarly, Beckett et al. (2017), using a principal component analysis (PCA) in a genotypic matrix composed of 77314 SNPs, clustered 283 ex-PVP lines and 66 key ancestors into three major groups: stiff stalk, Iodent, and non-stiff stalk heterotic groups. Nevertheless, PCA cluster the individuals based on the number of markers that are identical by state between them and not in alleles identical by descent. White et al. (2020), using a diversity estimate of 1 minus the probability of pairwise identity-by-state
between 329 ex-PVP and 8 ancestors, found eight heterotic subgroups represented by the lines B14, B37, B73, Oh43, PH207, PHZ51, Mo17, and PHGG7. As stated in the literature, DNA segments that are identical by descent are, by definition, identical by state (IBS), but not all IBS segments are copies coming from a common ancestor (Bernardo 2010; Falconer and Mackay 1996; Hartl et al. 1997; Walsh and Lynch 2018), and thus, estimating kinship from IBS segments could give an overestimated value of the genetic diversity between individuals.

The genetic contribution from a set of progenitors to its descendants was defined as the theoretical number of genes that trace back to those progenitors (Delannay et al. 1983; Mikel et al. 2010). Commonly genetic contribution analysis has been based on pedigree information. For instance, Mikel and Dudley (2006), formulate a grouping of inbred lines based on familial genetic backgrounds to understand the composition of ex-PVP germplasm released from 1980 to 2004, showing that Holden’s Foundation Seed company lines had a significant contribution of the lines Mo17 and LH51, while Dekalb relied primarily on the use of the hybrid PHI3737. Smith et al. (2004), using a coefficient of parentage (CP) computed from pedigree information identified around 60 ancestors involved in the derivation of 68 important Pioneer Hi-Bred International (PHI) hybrids released from 1934 to 1997, concluding that only six ancestors (A237, AFLF, BLACKSDC0, Iowa Stiff Stalk Synthetic (BSSS), Golden Glow, and Krug) had mean contribution per era bigger than 5%. Mikel (2008), using CP concluded that the public line B73 has contributed significantly to the germplasm of 470 lines released from 1996 to 2005. Similarly, Mikel (2011), using KIN software (Tinker and Mather 1993) to estimate CP among 344 proprietary inbred lines released from 2004 to 2008, concluded that within the stiff stalk heterotic group the line B73 had a genetic contribution around 11.7% to newly developed germplasm, while B14 and B37 had a contribution of 3.2 and 1.5% respectively. Additionally,
Mikel (2011), found that the public line Mo17 contributes approximately 2% to the germplasm of 344 proprietary lines, with a higher contribution to lines released by the Monsanto breeding program (Mikel 2011). The use of pedigree information (i.e. CP) or IBS segments in the estimation of the genetic contribution of ancestors or the quantification of genetic diversity between populations has represented a way to better utilize individuals in breeding programs and to identify the best combination of parents to maximize agronomic performance (Beckett et al. 2017; Zhang et al. 2018). Most studies using pedigree-based or IBS approaches have clustered ex-PVP lines within one of the three major heterotic groups (stiff stalk, Iodent, and non-stiff stalk) described in the U.S. Corn Belt, with some of them providing a clear differentiation within each group (Beckett et al. 2017; Coffman et al. 2019; Mikel 2008, 2011; Mikel and Dudley 2006; Nelson et al. 2008b; Schaefer and Bernardo 2013; Smith et al. 2004; White et al. 2020). However, the authors are unaware of any experiment examining genetic diversity and/or ancestors’ genetic contribution in maize using kinship coefficients computed from estimates of identity by descent (IBD) segments. Therefore, the main purpose of this study was i) to quantify the size of IBD segments, ii) to estimate the number of generations from a common ancestor for each segment, iii) to quantify the genetic diversity of 44 ex-PVP and eight key ancestor lines when computed from IBD segments, iv) to compare diversity estimates calculated from pedigree information, IBS probabilities, and IBD segments, and v) to quantify the genetic contribution of the eight key ancestors.

Materials and methods

Plant material

A total of 44 ex-PVP inbred lines included in the factorial experiment in the Genomes to Fields initiative (https://www.genomes2fields.org/about/project-overview/#standards-and-methods) were selected (Table 3.1). From these lines, 11 had a stiff stalk background (SSS),
while the other 33 had a non-stiff stalk (NSS) background (Table 3.1). Pedigree information for each of the 44 inbred lines was obtained from U.S. PVP certificates, U.S. utility patent supporting information, and registration articles. Eight inbred lines, B14, B37, B73, PHJ40, PHG39, Mo17, Oh43, and PH207 were selected as key ancestor lines. These lines were selected based on the number of times each was a parent of other PVP inbreds (Mikel and Dudley 2006; van Heerwaarden et al. 2012). Key ancestor lines B14, B37, B73, PHJ40, and PHG39 were selected as key lines within the stiff stalk heterotic group, while the remaining were classified as key lines within the non-stiff stalk heterotic group (Mikel and Dudley 2006).

**Integration of genotypic datasets**

Imputed genotypic data, based on B73 AGPv4 coordinates, was obtained from the Genomes to Fields (G2F) initiative, field season 2017, and from Panzea ZeaGBSv2.7 (www.panzea.org). Both genotypic databases are publicly available to download through the CyVerse Data Store. The G2F database contained 945574 SNPs and 4402 samples, while the Panzea ZeaGBSv2.7 database contained 955690 SNPs and 17280 samples. To use information from both datasets, 943103 common SNPs were identified and retained in a merged consensus SNP panel. A subset of 52 samples from the consensus SNP panel, representing the 44 inbred lines and the 8 lines used as key ancestors, were extracted. Out of the 52 unique lines, 39 were found in both datasets and 13 only in the G2F dataset. A test for the differences between the 39 common inbred lines was performed by looking for mismatches between both datasets. All inbred lines had a mismatching rate between 0 and 2.4% (data not shown). Sample filtering was done by keeping those with the lowest level of heterozygosity. After the quality control process, the genetic information of 20 inbred lines was obtained from the Panzea ZeaGBSv2.7 (www.panzea.org), while 32 were obtained from the G2F genotypic dataset (Table 3.1).
SNP calling

Marker filtering was performed using PLINK 1.9 (Chang et al. 2015) and TASSEL version 5.2.52 (Bradbury et al. 2007). A total of 23829 markers with insertions and/or deletions were removed from the analysis. Multiallelic sites were removed by using the option “--biallelic-only strict” in Plink 1.9 (Chang et al. 2015). SNPs with rates of missing information above 5% were removed to ensure genotype concordance and to exclude poor performing markers. Monomorphic markers among the 52 pedigrees were removed. Genotypes were phased and imputed by using Beagle version 5.1 package (Browning et al. 2018). Physical distances for each marker were converted to genetic distances by using a dense 0.2 cM resolution map created by Ogut et al. (2015), with a genetic distance of 1385.608 Kb per cM. They create the linkage map by using 7386 SNP markers from 25 families (4421 genotyped individuals) selected from the nested association mapping (NAM) population (Ogut et al. 2015). After completing data-cleaning, the genotypic data contains 160158 sites for each of the 52 lines, with coverage of 2102.757 Mb (1517.57 cM) of the genome and with a marker every 13.36 Kb on average.

IBD segments and generations from a common ancestor

Regions in the genome that have been inherited from common ancestors (IBD segments) were identified with the identity by descent linkage disequilibrium (IBDLD) program version 3.38 (Han and Abney 2011, 2013). IBDLD program uses a probabilistic approach with a hidden Markov model to estimate IBD segments in pairs of individuals (Han and Abney 2011). The HMM requires the specification of three probabilities: prior, transition, and emission (Han and Abney 2011; Westhead and Vijayabaskar 2017; Zucchini et al. 2017). The prior probability describes the probability of IBD for the first marker and is assumed to be the condensed identity coefficient for that SNP (Han and Abney 2011). The transition probability describes the likelihood of transitioning from the IBD status of marker \( k - 1 \) to marker \( k \) (Han and Abney...
IBD transition probability depends on the distance between markers, and thus, variations in the rate of recombination are included in the model (Browning and Browning 2012; Leutenegger et al. 2003). Finally, the emission probability describes the likelihood of the observed genotype at maker $k$ given the unknown hidden state of the marker (Westhead and Vijayabaskar 2017). IBDLD program further expresses the emission probability conditioned on the true genotype of $n$ previous loci to account for linkage disequilibrium (LD) (Han and Abney 2011). Since the true genotype at a locus is not known, the model allows the incorporation of genotyping errors ($\varepsilon$), such that the HMM estimates the IBD probabilities at each SNP conditional not only on multilocus genotypic information and LD but on genotyping errors and/or missing data (Han and Abney 2011). The model uses ridge regression to calculate LD patterns by conditioning the genotypes of $n$ previous SNPs ($n = 10$ for this experiment) as well as the genotypes of all previous SNPs within $k$ cM ($k = 2$ for this experiment). The method “GIBDLD” within the IBDLD program was used. The GIBDLD method uses a modified HMM to avoid pedigree information in the computation of the IBD probabilities (i.e. prior probability) and to enable the use of related or unrelated individuals (Han and Abney 2013). IBD segments, for each pair of individuals, were constrained to have a minimum length of 500 Kb, have more than 50 SNPs, and have SNPs with an IBD probability above 70%. These requirements force the segment to be homozygous for long sections, avoiding segments formed by chance (McQuillan et al. 2008). A histogram for the length of each IBD segment in a logarithmic scale ($\log_{10}[\text{IBD length in Mb}]$) for easy visualization was produced for each heterotic group. According to Fisher (1954) and Stam (1980), the expected length of an IBD segment in the genome follows an exponential distribution with a mean equal to $(2g)^{1}$ Morgans, where $g$ corresponds to the number of generations from a common ancestor. Therefore, the expected number of generations ($g$) from
a common ancestor was calculated by the inverse of twice the length (L), in Morgans, of each segment \((g=(2L)^{-1})\) (Fisher 1954; Howrigan et al. 2011; Stam 1980; Thompson 2013).

**Genetic diversity among ex-PVP lines and dendrogram**

The degree of relationship among the 44 ex-PVP inbred lines and the eight key ancestors, was estimated using three approaches. The first consists of using the posterior IBD probability estimated with the IBDLD program to compute an empirical genomic relationship matrix (marker-based kinship). Briefly, the program computes the kinship coefficient as

\[
\hat{\pi}_i = \Delta_{1,i} + \frac{1}{2} (\Delta_{3,i} + \Delta_{5,i} + \Delta_{7,i}) + \frac{1}{4} \Delta_{8,i},
\]

where \(\Delta_{r,i}\) is the posterior probability for the condensed identity coefficient \(r (r =1, 3, 5, 7, \text{ or } 8)\) at locus \(i\) (Han and Abney 2013; Jacquard 1974). According to Jacquard (1974), the condensed identity coefficients \(\Delta_{1,i}, \Delta_{3,i}, \Delta_{5,i}, \Delta_{7,i}\) and \(\Delta_{8,i}\) correspond to cases in which the number of alleles that are IBD between two diploid individuals is 4, 2, 2, 2, and 1 respectively. Han and Abney (2011), calculated, at a genome-wide scale, the kinship coefficient for the pair of individuals as

\[
\pi = M^{-1} \sum_{i}^{M} \hat{\pi}_i,
\]

where \(M\) corresponds to the total number of markers used (Han and Abney 2013). The second approach the marker data to estimate an IBS marker-based kinship coefficient (G kinship) computed with the AGHmatrix package (Amadeu et al. 2016). In the present study, the G kinship was computed using VanRaden (2008) methodology, by solving

\[
K = \frac{ZZ'}{2\Sigma p_i(1-p_i)},
\]

where \(Z\) correspond to a matrix product of the difference between the maker matrix \(X_{[n \times m]}\) (with \(n\) individuals and \(m\) markers) from a matrix in which the columns corresponds to \(2(p_i - 0.5)\), where \(p_i\) was the frequency of the second allele at marker \(i\). The denominator in \(K\), was used to scale the matrix such that this becomes analogous to the additive relationship (VanRaden 2008). The third approach was a pedigree-based kinship estimated from pedigree information collected for each inbred line. Pedigree-based kinship coefficients were computed with the AGHmatrix package (Amadeu et al.
2016), using the recursive method proposed by Henderson (1976), where the diagonal correspond to \(1 + f_i\) and off-diagonal \(r_{ij}\sqrt{f_i f_j}\) for every \(f_i\) and \(f_j\) representing the inbreeding coefficient for individual \(i\) and \(j\) respectively, and \(r_{ij}\) the Wright’s coefficient of relationship between individuals (Crow and Kimura 1970; Henderson 1976; Wright 1922). Genetic diversity (GD) for each inbred among the 52 lines (44 ex-PVP and 8 key ancestor lines) was calculated by generating a distance matrix in which each element was calculated as 1 minus the kinship coefficient. Therefore, three kinship distance matrices, marker-based, G-based, and pedigree-based, were created. Both the marker-based and G-based distance matrices measures the genetic diversity computed from molecular markers, however, the marker-based uses the posterior probability of the IBD status at each locus, while the G-based distance uses estimates of identity-by-state (IBS) in the calculation. All distance matrices were used to create a dendrogram applying the unweighted pair group method with arithmetic mean (UPGMA) within the “hclust” function in fastcluster package (Müllner 2013).

**Key ancestor lines genetic contribution**

The genetic contribution of the eight key ancestor lines to each of the 44 ex-PVP inbreds was estimated for both the pedigree-based and the marker-based kinship. The kinship for each key ancestor with the ex-PVP lines was used as an estimate of the relative genetic contribution the key ancestor had to the genetic constitution of every ex-PVP line. The relative contribution of each ancestor to all lines within a heterotic group was considered as the mean genetic contribution of the ancestor to the heterotic group. Estimation of IBD segments with genotypic data allows the quantification of the proportion of the covered genome descended from each key ancestor. Therefore, the number of times an SNP was included in a shared IBD segment between the ex-PVP and the key ancestor lines was estimated. SNPs not included in an IBD segment were
referred to as non-IBD markers, while those within an IBD segment were labeled with the key ancestor to which it is sharing the segment. Then, the proportion of the genome descended from the key ancestor was calculated by dividing the total number of SNPs classified as identical by descent by the total number of polymorphic SNPs used (160158 sites).

Results

Genetic diversity among ex-PVP lines and dendrogram

Genetic diversity (1 – kinship) among the ex-PVP lines from the NSS heterotic group was greater than diversity among SSS heterotic group, with values of 0.92, 0.96, and 0.93 for the marker-based, G-based, and pedigree-based distance matrices respectively (Figure 3.1). Diversity within the SSS heterotic group (11 ex-PVP lines) was estimated at 0.81 and 0.83 when computed with the marker-based and G-based distance matrices, while it was 0.79 when computed with pedigree information (Figure 3.1). The diversity between groups (SSS vs NSS) was estimated at 0.97, 1.15, and 1.00 for the marker-based, G-based, and pedigree-based estimates respectively (Figure 3.1). The 52 inbred lines corresponding to 44 ex-PVP and 8 key ancestors, were grouped into 11 clusters when diversity was computed from IBD segments (Figure 3.2) or IBS segments (Figure 3.3), and in 19 clusters for the pedigree-based distance matrix (Figure 3.4). All diversity estimates subdivide the lines into the heterotic subgroups Mo17, Oh43, PH207, B14, B37, PHJ40, PHG39, B73, and Iodent Lancaster (representative line PHZ51). For the diversity estimated with the pedigree-based distance matrix, the lines Q381, PHW03, PHP60, PHM49, PHK76, PHJ65, PHHV4, and LH162, from which pedigree information was obtained only for one generational cycle (Table 3.1), were clustered in individual groups (Figure 3.4), while when the marker-based or G-based distance matrices were used, these lines were grouped into one of the nine heterotic subgroups with exception of the lines PHW03, PHK05, CG120, and CG123 (Figure 3.2 and 3.3). The relationship between the
ex-PVP lines and the key ancestors was greater for the SSS heterotic group for both kinship estimators (Table 3.2). G kinship matrix cannot be interpreted as a probability but as the correlation that 2 homologous genes between or within individuals are shared by descent, with values explaining the excess, for positive values, or deficit, for negative values, of IBD relative to the IBD obtained for alleles drawn at random from the reference population (Wang 2014). Within the B73 heterotic subgroup, all lines (LH195, LH198, and PHHV4) had more than 34 % of the covered genome in IBD with the B73 key ancestor (Figure 3.4). Inbred lines PHP38, PHJ65, and PHW52 had 31.4, 42.4, and 36.1 % of the genome, respectively, in IBD with the ancestor PHG39 (Figure 3.5). Inbred lines PHK05, PHW03, CG120, and CG123 were clustered near the SSS heterotic group branch (Figure 3.2), however, 55.0, 54.3, 89.7, and 75.0 % of the genome, respectively, corresponded to non-IBD markers with the eight key ancestors used (Figure 3.5). Similarly, the lines LH82, PHM57, and PHP60, clustered within the NSS heterotic group branch (Figure 3.2 and 3.3), had 62.6, 85.9, and 95.2 % of the genome corresponding to non-IBD SNPs with the key ancestors used in the experiment (Figure 3.5). All ex-PVP lines clustered in the PH207 heterotic subgroup had between 28.7 % (with line PHM49) and 90.1 % (with line Q381) of the genome in IBD with the PH207 key ancestor (Figure 3.5). Inbred line PHRE1 had 46.4 % of the genome in IBD with the key ancestor PHJ40, while the line PHK76 had a genome composed mainly with the ancestors PHJ40 (20.9 %), Oh43 (18.1 %), and Mo17 (15.4 %) (Figure 3.5). The proportion of loci in the PHJ65 inherited from the key ancestors of the NSS group was below 1.1%, while the lines B73 and PHG39 had 36.4 and 31.4 % of the genome in IBD with PHJ65 respectively (Figure 3.5).

**Key ancestor lines genetic contribution**

The mean genetic contribution of key ancestors to ex-PVP lines estimated with pedigree-based kinships was lower than with marker-based or G-based methods for both heterotic groups
(Table 3.2). In general, the ancestor PHJ40 had a lower genetic contribution to the 11 ex-PVP line members of the SSS heterotic group for both kinships, while B37 and B73 had the higher contribution when the pedigree-based kinship was used (Table 3.2). For the marker-based kinship, the ancestors B14, B37, B73, and PHG39 had a similar contribution to the lines within the SSS heterotic group (Table 3.2). The ancestor PH207 had the highest contribution to the ex-PVP lines of the NSS heterotic group with 0.16 for both kinship estimators (Table 3.2). Pedigree-based kinship coefficients indicate that the lines LH162, LH38, PHM49, PHK05, PHJ65, PHK76, PHP60, and Q381 were unrelated to the key ancestors (kinship of zero), while marker-based coefficients were able to detect some hidden relationships (data not shown). On average, inbred lines within the SSS background had 18.1 % of the genome classified as identical by descent with the ancestor B73, while the lines within the NSS background had 20.6 % with the line PH207 (data not shown). In general, the degree of relationship among the key ancestor lines representative of the SSS heterotic group (B14, B37, B73, PHG39, and PHJ40) was higher than the relationship between Mo17, Oh43, and PH207 (Table 3.2). Pedigree-based coefficients showed a kinship of 0.04 between B37 and PHG39 and B14 and PHG39, while marker-based coefficients found that these lines were more related to, with kinship of 0.34 and 0.35 respectively (Table 3.2). Finally, pedigree-based estimators suggest a kinship between lines B73 and Mo17 of 0.02, while marker-based estimators found a relationship of 0.03 between both lines (Table 3.2).

**IBD segments and generations from common ancestor**

A total of 160158 polymorphic SNPs covering 2102.757 Mb of the genome were used in the estimation of IBD segments among the 44 ex-PVP and 8 key ancestor lines. Key ancestor lines B14, B37, B73, PHG39, and PHJ40 were included as key lines within the stiff stalk (SSS) heterotic group, while the lines Mo17, Oh43, and PH207 were classified as key lines within the
non-stiff stalk (NSS) heterotic group (Mikel and Dudley 2006). Comparison between the ex-PVP lines and key ancestors within heterotic groups reveals that fifty percent of the identified IBD segments had a length below 4.8 Mb (Q2 = 0.68 log_{10}[Mb]) and 5.2 Mb (Q2 = 0.72 log_{10}[Mb]) for the NSS and SSS heterotic group respectively (Figure 3.6). Similarly, 25% of the identified segments had a mean length above 14.8 Mb (Q3 = 1.17 log_{10}[Mb]) and 13.8 Mb (Q3 = 1.14 log_{10}[Mb]) for the SSS and NSS heterotic groups respectively (Figure 3.6). For both heterotic groups, the IBD segments between the first (Q1) and third (Q3) quantile had a length between 0.68 and 1.17 log_{10}[Mb], which correspond to a length between 4.8 and 14.8 Mb (Figure 3.6). The mean length of IBD segments was slightly higher than the third quantile for both heterotic groups (Figure 3.6), revealing the prevalence of large IBD segments inherited from the key ancestors. In general, 16.1 and 17.5% of the IBD segments identified within the stiff stalk and non-stiff stalk heterotic group were generated in no more than 5 generations of recombination (Figure 3.7). For both heterotic groups, most of the segments corresponded to an inbreeding event formed within 5 to 15 generations ago, while around 25% of the IBD segments correspond to an inbreeding created between 35 and more than 135 generations of recombination from the common ancestor (Figure 3.7).

**Discussion**

**Genetic diversity among ex-PVP lines and dendrogram**

Genetic diversity estimates were higher for the inbred lines within the NSS compared to the SSS heterotic group (Figure 3.1). The relationship matrix computed from IBS, as the one computed with VanRanden’s (2008) methodology (G matrix), scales the relationship matrix to make it analogous to the pedigree-based kinship matrix, where the diagonal correspond to 1 + f_i and off-diagonal \( r_{ij} \sqrt{f_i f_j} \), where \( f_i \) and \( f_j \) represents the inbreeding coefficient for individual \( i \) and
\( j_i \) and \( r_{ij} \) the Wright’s coefficient of relationship between individuals (Crow and Kimura 1970; Henderson 1976; Wright 1922). Therefore, G-based kinship matrix has diagonal elements ranging between 1, for non-inbred individuals, and 2, for inbred lines (Endelman and Jannink 2012). For that reason, genetic diversity estimates between heterotic groups have values bigger than or equal to 1 when calculated with the G-based or pedigree-based distance matrices, while for the marker-based, the estimates correspond to the posterior probability of IBD at each locus and its maximum will be 1. Mikel (2008) in a study using kinship calculated from pedigree information in 470 ex-PVP lines released from 1996 to 2005, reported a similar behavior with genetic diversity of 0.64 and 0.90 within the stiff stalk and non-stiff stalk lines, respectively. Genetic diversity between groups was 0.97, 1.15, and 1.00 when computed with a marker-based, G-based, and pedigree-based distance matrix, respectively, revealing a greater diversity between groups than within heterotic groups. Commercial breeding programs around the world continue producing crosses between genetically diverse heterotic groups to maximize performance, however, each group is mainly bred by the recycling of closely related inbred lines (Mikel 2008), resulting in less genetic diversity within than between groups.

Genetic diversity (1 − kinship) was estimated for all kinship methods by computing a marker-based, G-based, and pedigree-based distance matrix. Marker-based and G-based distance matrices use genotypic data in the estimation of the probability of IBD or IBS between and within individuals, respectively. Both marker-based and G-based distance matrices separate the stiff stalk heterotic group into the subgroups associated with the inbred lines B14, B37, B73, PHG39, PHJ40, PHK05, PHW03, CG120, and CG123 (Figure 3.2 and 3.3). The lines PHK05, PHW03, CG120, and CG123 were clustered near the SSS heterotic group and therefore were classified as other SSS heterotic subgroups (Figure 3.2). The genetic similarity between the eight
key ancestors and the lines within the other SSS group was below 13.1\% (Figure 3.5), suggesting these lines were derived from other public or proprietary lines not accounted for in this experiment. White et al. (2020), using a neighbor-joining clustering analysis from a genetic distance matrix computed from identity by state across 337 ex-PVP inbreds, clustered the lines PHW03 and PHK05 near the Flint heterotic subgroup, an early-maturing group considered to act within the stiff stalk. Inbred CG123, developed in Guelph, Canada, was derived from the lines NK804 and NK848, both PHI3737 derivatives. Results for the CG120 and CG123 were inconsistent with the well-known performance of the stiff stalk heterotic group, however, information about the genetic contribution of the key ancestors indicated that CG120 had 13.1 and 11.1\% of the genome in IBD with the ancestors B14 and PH207, respectively, while CG123 had 11.3 and 7.8\% from B37 and PH207 respectively (Figure 3.5). Ex-PVP lines of the NSS heterotic group were divided into the subgroups Mo17, Oh43, PH207, Iodent Lancaster (representative line PHZ51), and other non-stiff stalks (Other NSS) when computed with the marker-based or G-based diversity (Figure 3.2). There was no difference in the clustering of the 52 inbred lines when comparing both marker-based and G-based diversity analysis, however, the first provides estimates of the probability of IBD at each locus, while the later uses the identity-by-state (IBS) at casual loci as an approximation to the probability of IBD (Yu et al. 2006). Diversity analysis using a pedigree-based distance matrix identifies the same main heterotic subgroups: B14, B37, B73, PHG39, PHJ40, Mo17, Oh43, PH207, and Iodent Lancaster (Figure 3.4). In the analysis of genetic diversity, the key ancestors B14, B37, B73, PHG39, PHJ40, Mo17, Oh43, and PH207 were included because each was known to be present in the pedigree of most of the lines used in this study and they have constituted some of the most recombined inbreds appearing in the derivation of other PVP lines in the U.S. (Mikel and Dudley 2006; van
Heerwaarden et al. 2012). Therefore, as expected, the main heterotic subgroups identified corresponds to groups associated with the eight key ancestor lines plus the line PHZ51 (for the Iodent Lancaster group).

When using pedigree information nine additional subgroups corresponding to the lines LH162, LH38, PHG83, PHHV4, PHJ65, PHK76, PHP60, PHW03, and Q381 were identified because of the lack of pedigree information (Figure 3.4), however, those lines were included in one of the main heterotic subgroups when the marker-based, computed from IBD posterior probabilities (Figure 3.2), or G-based, computed from IBS probabilities (Figure 3.7), were used. The analysis using the marker-based kinship coefficient to compute a distance matrix grouped some inbred lines close to certain ancestors. For instance, the proportion of loci in the line PHP38 inherited from the key ancestors PHG39 was 42.4 % (Figure 3.5), locating the line within the PHG39 heterotic subgroup (Figure 3.2). Mikel et al. (2010), using a dissimilarity matrix computed from pedigree-based kinship coefficients, clustered PHP38 inbred line within the “B73” group, while White et al. (2020), using kinship calculated from identity by state markers, clustered the line PHP38 within a heterotic subgroup with representative line PHG39. This PHG39 subgroup was considered to be bred by Pioneer Hi-Bred International and to be within the stiff stalk heterotic group (White et al. 2020). In general, the line PH207 had the highest proportion of the genome identical-by-descent with ex-PVP lines from the NSS group, with an average of 52.9% of the PH207 genome identical to lines in the PH207 heterotic subgroup (Figure 3.5). Although the selection of the eight key ancestor lines results in large segments shared with the ex-PVP, for the lines PHK05, PHM57, PHP60, PHW03, CG120, CG123, and LH82 the percentage of SNPs classified as non-IBD was above 54 % (Figure 3.5). Given that these lines were derived by Pioneer Hi-Bred International (PHI), Holden's Foundation Seeds, and
the University of Guelph, Canada (Table 3.1), there could be another public or private line(s) that could account for more shared SNPs than the ones reported here. The genetic diversity estimates used in this experiment rely on the estimation of kinship coefficients, which describes the relationship between or within individuals relative to an ancestral population (Falconer and Mackay 1996; Malécot 1948; Wright 1921, 1922). Finally, the proportion of the genome in IBD between the line PHJ65 and the key ancestors Mo17, Oh43, and PH207 was estimated to be lower than 1.1%, while it was 36.4% for the key ancestor PHG39 (Figure 3.5). Therefore, the ex-PVP line PHJ65 was clustered within the PHG39 heterotic subgroup (Figure 3.2). Based on pedigree information collected in the present study, the line PHJ65 was classified in the non-stiff stalk heterotic group, however, our analysis of IBD segments suggests a higher resemblance with the Iowa stiff stalk synthetic population derived progenitors. Close attention should be paid to PHJ65 because other studies (using IBS estimates) classified this line within the NSS heterotic group. For instance, White et al. (2020), clustered PHJ65 inbred within the Iodent Lancaster subgroup (representative line PHZ51), while Beckett et al. (2017) using principal components grouped the line within the non-stiff stalk, near to the line PHP60. Differences in our results with those reported by White et al. (2020) and Beckett et al. (2017) could rely on either the lack of other public and/or proprietary lines like PH595, PHG47, PH814, PH848, and PHP02 among others, that are known because of their significant contributions in the development of important lines within Pioneer Hi-Bred International breeding programs (Mikel 2011; White et al. 2020) and/or on the use of a genome-wide estimate of IBD instead of genome-wide estimate of identity by state.

**Key ancestor lines genetic contribution**

On average, there was a similar genetic contribution of the key ancestors B14, B37, B73, and PHG39 to the lines of the SSS heterotic group when estimated with the marker-based
(approximately 23%), while the contribution of the lines B37 and B73 was higher (approximately 33%) when computed with pedigree-based kinships (Table 3.2). The key ancestor PHJ40 had the lowest genetic contribution to the ex-PVP lines of the SSS heterotic group, with 10 and 11% for the marker- and pedigree-based estimators respectively (Table 3.2). For the NSS heterotic group, the genetic contribution of the line PH207 was 16% when computed with both pedigree- and marker-based kinships (Table 3.2). A high correlation (r = 0.86) between kinship estimators was observed (data not shown), suggesting that relationships caused by recent ancestry have the biggest contribution to the variability of the coancestry between individuals because old ancestry contributes uniformly to the kinship in the sample of genotyped individuals (Toro et al. 2002). Differences in the estimators of coancestry between the pedigree-based and marker-based kinships could be explained as a consequence of some population-specific parameters and to some variations in the estimation of the relationship. The effective population size, selection, and the random Mendelian sampling could be included as some population-specific parameters that will affect both types of coefficients, while pedigree depth and the number of markers correspond to restrictions affecting directly how close the estimator will be to the true value. The effective population size (\(N_e\)) has a direct influence in the levels of linkage disequilibrium (LD) in a population (Crow and Kimura 1970; Hill and Robertson 1968; Hill 1981; Laurie-Ahlberg and Weir 1979), with lower LD associated with larger \(N_e\) (Sved 1968; Wang 2016). Similarly, high recombination levels reduce the amount of disequilibrium between loci, with a reduction in LD of one half per generation when the loci are unlinked (Bulmer 1971). Consequently, LD will increase the probability of finding blocks of loci where all alleles are either identical or non-identical by descent, affecting kinship when computed with pedigree or marker data (Sved 1968; Wang 2016; Weir et al. 1980). According to
Wang (2016), an increase in the effective population size will result in decreased LD and therefore, a large number of markers will be required to capture the relationship between individuals. Similarly, selection changes the allele frequencies in the selected offspring because only individuals with superior performance for the selected trait will contribute with alleles to the next generation (Albrechtsen et al. 2010). Therefore, if an allele affects a trait under selection, selection will increase the probability that multiple individuals inherit identical alleles. Finally, because of random Mendelian sampling, two individuals with the same pedigree will have different marker-based kinships, while the two individuals will have the same kinship when computed with pedigree information. Therefore, the ability of marker data to account for changes in allele frequencies, for mendelian sampling and for the effect of selection on mendelian sampling results in differences between marker- and pedigree-based kinship.

Pedigree-based kinship coefficients provide the probability of alleles, at a locus, being identical by descent relative to known common ancestors in a declared reference population. The declared reference population is a function of the depth of the pedigree, i.e. on the inclusion or exclusion of common ancestors in the pedigree. In contrast, the marker-based kinship coefficient relies on the assumption that marker allele frequencies in the reference population are known (Goudet et al. 2018; Wang 2014). Generally, allele frequencies in the reference population are not known, and therefore, with marker-based kinship, the sample of individuals genotyped is used to estimate allele frequency in the reference population (Wang 2014). As a consequence, marker-based kinship cannot be interpreted as a probability but as the correlation that 2 homologous genes between or within individuals are shared by descent, with values explaining the excess or deficit (for negative values) of IBD relative to the IBD obtained for alleles drawn at random from the reference population (Wang 2014). Therefore, if the sample of genotyped
individuals constitutes a representative and random sample of the reference population, kinship values computed from marker data will be unbiased estimates of kinship. Additionally, Engelsma et al. (2012), states that the correlation between pedigree and marker-based estimators decreases as the number of markers decreases. Wang (2016) concluded that an improvement in the kinship estimates is obtained when many markers (> $10^4$ SNPs) are used. Nevertheless, when good and ample pedigree information is available for the analysis (i.e. good pedigree depth), pedigree-based estimates will be close to marker-based estimates, however, a difference will persist because the inability of the pedigree-based estimators to capture the mendelian sampling. In general, both pedigree- and marker-based kinship coefficients rely on the definition of a reference population, however, the latter can shift the reference to be the population of sampled individuals. Therefore, with a large number of markers and a good sample of individuals representing the current population, marker-based coefficients produce precise and unbiased kinships avoiding the problem of lack of pedigree information. Henceforth, a probabilistic model that uses a hidden Markov model (HMM) method to identify IBD segments between a pair of individuals could reveal regions in the genome that were not possible to identify with a pedigree-based analysis and thus, should be the preferred methodology to study genetic diversity when marker information is available.

**IBD segments and generations from common ancestor**

Our results provide a high-resolution data analysis in the identification of regions in the genome inherited from the key ancestors representative of the stiff stalk (B14, B37, B73, PHG39, and PHJ40), and non-stiff stalk (Mo17, Oh43, and PH207) heterotic groups. Overall, there was a prevalence of large IBD segments (with a mean length of 15.5 and 16.2 Mb for the SSS and NSS heterotic groups respectively) inherited from the key ancestor lines (Figure 3.6), with a trend of decreasing the number of segments as the length of the segment increases. Large
preserved regions in the genome could be associated with selection processes, resulting in long
DNA segments inherited as a block from the parents (Hill and Robertson 1968; Jordan et al.
2004; Robertson 1961; Thompson 2013). Therefore, under positive selection, that is, a selection
favoring a phenotype, a slight increase in LD surrounding the favored allele will be produced
(Albrechtsen et al. 2010; Hospital and Chevalet 1993). In these cases, the length of the IBD
segment surrounding the alleles subject to selection will increase, experiencing less
recombination at the population level (Albrechtsen et al. 2010). Golding and Strobeck (1980)
and Albrechtsen et al. (2010), state that a reduced recombination rate in the genome, leading to
significant LD, could be explained as a function of the effective population size. These could
partially be explained by an increase in random genetic drift because of the small $N_e$, which in
turn will increase the length of DNA that will be shared between individuals in the population
(Albrechtsen et al. 2010). Albrechtsen et al. (2010), concluded that the detection of long IBD
segments in populations could be used as evidence for strong and recent selection processes
because these segments have not suffered from recombination (or mutation). In cases where
alleles within long IBD segments are in linkage disequilibrium, specifically in a repulsion phase,
unfavorable alleles will persist in the population inducing the hitchhiking effect and reducing the
genetic diversity (Hospital and Chevalet 1993). This hitchhiking will create an increase in
genetic drift and a significant decrease in the effective population size (Barton 2000; Smith and
Haigh 1974). Therefore, the relationship between long IBD segments and effective population
size could be viewed as an outcome of selection in populations with small $N_e$. Palamara et al.
(2012), concluded that long IBD segments are common in populations with small $N_e$ because a
lower rate of coalescent events. The higher lengths of IBD segments found in the SSS
background are evidence of small $N_e$ and strong or recent selection for this group, which has
relied mainly on Iowa Stiff Stalk Synthetic derived lines for the accumulation of favorable alleles. Our results reflect the maize breeding process in the U.S., which is known for the use of a small subset of elite inbred lines and for the recycling of closely related lines within breeding pools to derive new germplasm (Mikel 2011).

The number of generations of recombination from a common ancestor ranged between 1 to more than 135 generations, with most of the IBD segments corresponding to an inbreeding created in less than 15 generations (Figure 3.7). Following Stam (1980) and Fisher (1954), 25 generations of recombination correspond to shared IBD segments of 2 cM (2.8 Mb) on average. For the NSS and SSS heterotic groups, around 62.5 and 64.4% of the IBD segments (Figure 3.7), have a length above 2.8 Mb. Small shared IBD segments, corresponding to an inbreeding created between 35 and 135 generations, account for around a quarter of the total IBD segments identified in both heterotic groups (Figure 3.7). Higher number of generations reduce the size of the segments because of several cycles of recombination. According with Browning and Browning (2013), HMM probabilistic methods are known to be precise and accurate in the detection of large shared IBD segments (commonly > 2 cM), while for small segments these models have reduced power, resulting in the overestimation of the ends of an IBD segment and in the underestimation of segment length for contiguous small segments. Large IBD segments shared between a set of ex-PVP and the six ancestor lines allows the estimation of genetic diversity and ancestor’s genetic contribution to new released lines. Accurate and improved detection of short IBD segments is highly used in analysis aimed to IBD mapping, haplotype phasing, and/or to improve population structure when unrelated individuals are used (Browning and Browning 2013; Browning and Browning 2012; Kong et al. 2009; Palamara et al. 2012), while long segments are used commonly for the estimation of relatedness between individuals.
According to Browning and Browning (2013) and Howrigan et al. (2011), long segments shared by descent are indicative of intensive selection pressures and recent relatedness, while short segments provide information about genetic bottlenecks or past consanguinity in the population. Finally, because artificial selection applied in plant breeding programs, reducing the number of effective mates ($N_e$), longer IBD segments, created no more than 25 generations, are expected, resulting in high kinships and long IBD segments distributed across the genome in high frequencies (Curik et al. 2014; Ferenčaković et al. 2013; Rebelato and Caetano 2018).

**Summary**

The use of a probabilistic approach to identify IBD segments between a set of progenitors and offspring, with known or unknown pedigree information, could not only help in understanding the genetic diversity of the germplasm but in the identification of line derivation processes, in the utilization of lines in breeding programs, and/or to identify the best combination of inbreds to maximize hybrid performance. Our results showed that both marker-based and G-based distance matrices provides the same results, however the marker-based provides actual IBD probabilities and not an approximation from IBS. Results obtained in the present study provide a high-resolution data analysis that allows the identification of genome wide IBD segments for the 44 ex-PVP inbred lines and the eight key ancestors. In general, long IBD segments were identified between each key ancestor and ex-PVP. Long IBD segments between key ancestors and ex-PVP lines suggests a narrow genetic diversity and/or small effective population size ($N_e$) for both heterotic groups. For both heterotic groups, most of the segments corresponded to an inbreeding event formed no more than 15 generations ago, while around 25% of the IBD segments correspond to an inbreeding created in more than 35 generations of recombination. Genetic diversity within the inbred lines from the NSS background was higher than within the SSS background. This reduced variability within heterotic groups reveals the
recycling of inbred lines in the maize breeding strategies in U.S. Corn Belt. Marker-based
distance matrix, computed from the posterior probability of the IBD status of each marker,
separates the 44 ex-PVP lines into the heterotic subgroups B14, B37, B73, PHG39, PHJ40, Other
SSS, Mo17, Oh43, PH207, Iodent Lancaster (representative line PHZ51), and other NSS.
Diversity analysis using a pedigree-based distance matrix identifies the heterotic subgroups B14,
B37, B73, PHG39, PHJ40, Mo17, Oh43, PH207, and Iodent Lancaster. In cases, where pedigree
information is not available for a set of individuals, pedigree-based distance matrices set the
individual as a separate cluster. In that regard, the inbred lines LH162, LH38, PHG83, PHHV4,
PHJ65, PHK76, PHP60, PHW03, and Q381, from which pedigree information was obtained only
for one generation, form a separate cluster when using pedigree-based distance matrices,
however, when kinship from IBD segments is used, they were clustered in one of the main
heterotic subgroups identified. There was a high correlation between marker- and pedigree-based
distance matrices, with the former being able to capture distant relatedness, selection processes,
and mendelian sampling more accurately. There was a higher genetic contribution of the key
ancestors B14, B37, B73, and PHG39 to the 11 ex-PVP lines members of the SSS heterotic
group, compared to the genetic contribution of the lines Mo17, Oh43, and PH207 in the NSS.
Using marker-based estimators we found that a small group of ancestors (B14, B37, B73,
PHG39, PHJ40, Mo17, Oh43, and PH207) have contributed large proportions of the genome to
important PVP lines, with ex-PVPs from the SSS group having longer regions in the genome
inherited from a very small number of key ancestors. Finally, modern maize inbred lines derive
much of their genomes from a few set of key ancestors.
Tables and figures

Table 3.1. List of the 44 ex-PVP and 8 key ancestors lines evaluated, with genetic background, patent applicant company or institution, patent number, number of generations the pedigree is available (Generations), and dataset from where genotypic data was collected.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>B14</td>
<td>Key Ancestor</td>
<td>Iowa Stiff Stalk Synthetic C0</td>
<td></td>
<td>6</td>
<td>Panzea</td>
</tr>
<tr>
<td>B37</td>
<td>Key Ancestor</td>
<td>Iowa Stiff Stalk Synthetic C0</td>
<td></td>
<td>8</td>
<td>Panzea</td>
</tr>
<tr>
<td>B73</td>
<td>Key Ancestor</td>
<td>Iowa Stiff Stalk Synthetic C5</td>
<td></td>
<td>8</td>
<td>Panzea</td>
</tr>
<tr>
<td>Mo17</td>
<td>Key Ancestor</td>
<td>Lancaster Sure Crop</td>
<td></td>
<td>1</td>
<td>Panzea</td>
</tr>
<tr>
<td>Oh43</td>
<td>Key Ancestor</td>
<td>Lancaster Sure Crop</td>
<td></td>
<td>2</td>
<td>Panzea</td>
</tr>
<tr>
<td>PH207</td>
<td>Key Ancestor</td>
<td>Pioneer Hi-Bred International</td>
<td>8300144</td>
<td>8</td>
<td>G2F</td>
</tr>
<tr>
<td>PHG39</td>
<td>Key Ancestor</td>
<td>Pioneer Hi-Bred International</td>
<td>8300115</td>
<td>4</td>
<td>Panzea</td>
</tr>
<tr>
<td>PHJ40</td>
<td>Ancestor</td>
<td>Pioneer Hi-Bred International</td>
<td>8600133</td>
<td>4</td>
<td>Panzea</td>
</tr>
<tr>
<td>3IH6</td>
<td>NSS</td>
<td>Dekalb Plant Genetics</td>
<td>9300087</td>
<td>14</td>
<td>G2F</td>
</tr>
<tr>
<td>CG123</td>
<td>NSS</td>
<td>University of Guelph, Canada</td>
<td>Canada</td>
<td>8</td>
<td>G2F</td>
</tr>
<tr>
<td>CG124</td>
<td>NSS</td>
<td>University of Guelph, Canada</td>
<td>Canada</td>
<td>11</td>
<td>G2F</td>
</tr>
<tr>
<td>LH123HT</td>
<td>NSS</td>
<td>Holden's Foundation Seeds</td>
<td>8400030</td>
<td>5</td>
<td>Panzea</td>
</tr>
<tr>
<td>LH162</td>
<td>NSS</td>
<td>Holden's Foundation Seeds</td>
<td>9000123</td>
<td>1</td>
<td>G2F</td>
</tr>
<tr>
<td>LH185</td>
<td>NSS</td>
<td>Holden's Foundation Seeds</td>
<td>9400036</td>
<td>6</td>
<td>G2F</td>
</tr>
<tr>
<td>LH210</td>
<td>NSS</td>
<td>Holden's Foundation Seeds</td>
<td>9000050</td>
<td>3</td>
<td>G2F</td>
</tr>
<tr>
<td>LH38</td>
<td>NSS</td>
<td>Monsanto Technology, LLC</td>
<td>8000066</td>
<td>1</td>
<td>G2F</td>
</tr>
<tr>
<td>LH51</td>
<td>NSS</td>
<td>Holden's Foundation Seeds</td>
<td>8200062</td>
<td>2</td>
<td>Panzea</td>
</tr>
<tr>
<td>LH82</td>
<td>NSS</td>
<td>Holden's Foundation Seeds</td>
<td>8500037</td>
<td>5</td>
<td>G2F</td>
</tr>
<tr>
<td>PHG29</td>
<td>NSS</td>
<td>Pioneer Hi-Bred International</td>
<td>8600047</td>
<td>13</td>
<td>G2F</td>
</tr>
<tr>
<td>PHG47</td>
<td>NSS</td>
<td>Pioneer Hi-Bred International</td>
<td>8600131</td>
<td>6</td>
<td>G2F</td>
</tr>
<tr>
<td>PHG83</td>
<td>NSS</td>
<td>Pioneer Hi-Bred International</td>
<td>8500152</td>
<td>10</td>
<td>G2F</td>
</tr>
<tr>
<td>PHJ65</td>
<td>NSS</td>
<td>Pioneer Hi-Bred International</td>
<td>9000245</td>
<td>1</td>
<td>Panzea</td>
</tr>
<tr>
<td>PHJ89</td>
<td>NSS</td>
<td>Pioneer Hi-Bred International</td>
<td>9100092</td>
<td>7</td>
<td>Panzea</td>
</tr>
<tr>
<td>PHK05</td>
<td>NSS</td>
<td>Pioneer Hi-Bred International</td>
<td>8800001</td>
<td>5</td>
<td>Panzea</td>
</tr>
<tr>
<td>PHK56</td>
<td>NSS</td>
<td>Pioneer Hi-Bred International</td>
<td>9000247</td>
<td>7</td>
<td>G2F</td>
</tr>
<tr>
<td>PHK76</td>
<td>NSS</td>
<td>Pioneer Hi-Bred International</td>
<td>8800036</td>
<td>1</td>
<td>G2F</td>
</tr>
<tr>
<td>PHM49</td>
<td>NSS</td>
<td>Pioneer Hi-Bred International</td>
<td>8800211</td>
<td>2</td>
<td>G2F</td>
</tr>
<tr>
<td>PHM57</td>
<td>NSS</td>
<td>Pioneer Hi-Bred International</td>
<td>8900313</td>
<td>4</td>
<td>G2F</td>
</tr>
<tr>
<td>PHN37</td>
<td>NSS</td>
<td>Pioneer Hi-Bred International</td>
<td>8900315</td>
<td>4</td>
<td>Panzea</td>
</tr>
<tr>
<td>PHN47</td>
<td>NSS</td>
<td>Pioneer Hi-Bred International</td>
<td>8800217</td>
<td>9</td>
<td>G2F</td>
</tr>
<tr>
<td>PHN82</td>
<td>NSS</td>
<td>Pioneer Hi-Bred International</td>
<td>8900317</td>
<td>13</td>
<td>G2F</td>
</tr>
<tr>
<td>PHP60</td>
<td>NSS</td>
<td>Pioneer Hi-Bred International</td>
<td>8900319</td>
<td>1</td>
<td>Panzea</td>
</tr>
<tr>
<td>PHR03</td>
<td>NSS</td>
<td>Pioneer Hi-Bred International</td>
<td>9100097</td>
<td>7</td>
<td>G2F</td>
</tr>
<tr>
<td>PHR25</td>
<td>NSS</td>
<td>Pioneer Hi-Bred International</td>
<td>8800002</td>
<td>9</td>
<td>G2F</td>
</tr>
<tr>
<td>PHR55</td>
<td>NSS</td>
<td>Pioneer Hi-Bred International</td>
<td>9100098</td>
<td>7</td>
<td>Panzea</td>
</tr>
<tr>
<td>PHR63</td>
<td>NSS</td>
<td>Pioneer Hi-Bred International</td>
<td>8900321</td>
<td>13</td>
<td>Panzea</td>
</tr>
</tbody>
</table>
Table 3.1 Continued

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>PHW03</td>
<td>NSS</td>
<td>Pioneer Hi-Bred International</td>
<td>8900324</td>
<td>1</td>
<td>Panzea</td>
</tr>
<tr>
<td>PHW30</td>
<td>NSS</td>
<td>Pioneer Hi-Bred International</td>
<td>9100102</td>
<td>10</td>
<td>G2F</td>
</tr>
<tr>
<td>PHW53</td>
<td>NSS</td>
<td>Pioneer Hi-Bred International</td>
<td>9300116</td>
<td>10</td>
<td>G2F</td>
</tr>
<tr>
<td>PHZ51</td>
<td>NSS</td>
<td>Pioneer Hi-Bred International</td>
<td>8600132</td>
<td>6</td>
<td>Panzea</td>
</tr>
<tr>
<td>Q381</td>
<td>NSS</td>
<td>Quality Research Associates</td>
<td>8500098</td>
<td>1</td>
<td>G2F</td>
</tr>
<tr>
<td>CG120</td>
<td>SSS</td>
<td>University of Guelph, Canada</td>
<td>Canada</td>
<td>9</td>
<td>G2F</td>
</tr>
<tr>
<td>LH145</td>
<td>SSS</td>
<td>Holden's Foundation Seeds</td>
<td>8300102</td>
<td>6</td>
<td>G2F</td>
</tr>
<tr>
<td>LH195</td>
<td>SSS</td>
<td>Holden's Foundation Seeds</td>
<td>9000047</td>
<td>6</td>
<td>Panzea</td>
</tr>
<tr>
<td>LH198</td>
<td>SSS</td>
<td>Holden's Foundation Seeds</td>
<td>9200021</td>
<td>7</td>
<td>G2F</td>
</tr>
<tr>
<td>PHB47</td>
<td>SSS</td>
<td>Pioneer Hi-Bred International</td>
<td>8300141</td>
<td>6</td>
<td>G2F</td>
</tr>
<tr>
<td>PHHV4</td>
<td>SSS</td>
<td>Pioneer Hi-Bred International</td>
<td>9200084</td>
<td>1</td>
<td>G2F</td>
</tr>
<tr>
<td>PHN66</td>
<td>SSS</td>
<td>Pioneer Hi-Bred International</td>
<td>9100096</td>
<td>7</td>
<td>G2F</td>
</tr>
<tr>
<td>PHP38</td>
<td>SSS</td>
<td>Pioneer Hi-Bred International</td>
<td>9000250</td>
<td>6</td>
<td>Panzea</td>
</tr>
<tr>
<td>PHRE1</td>
<td>SSS</td>
<td>Pioneer Hi-Bred International</td>
<td>9300114</td>
<td>6</td>
<td>G2F</td>
</tr>
<tr>
<td>PHV63</td>
<td>SSS</td>
<td>Pioneer Hi-Bred International</td>
<td>8800039</td>
<td>7</td>
<td>G2F</td>
</tr>
<tr>
<td>PHW52</td>
<td>SSS</td>
<td>Pioneer Hi-Bred International</td>
<td>8800215</td>
<td>5</td>
<td>Panzea</td>
</tr>
</tbody>
</table>

NSS: non-stiff stalk heterotic group; SSS: stiff stalk heterotic group; †: Genotypic data obtained from the Genomes to fields initiative, field season 2017 (G2F) or from Panzea organization ZeaGBSv2.7.
Table 3.2. Degree of relationship between each key ancestor and ex-PVP per heterotic group. The kinship was calculated with a marker-based (an HMM framework for IBD), G-based (IBS kinship computed with VanRaden (2008) algorithm), and a pedigree-based kinship method.

<table>
<thead>
<tr>
<th>Key Ancestor</th>
<th>Heterotic Group</th>
<th>B14</th>
<th>B37</th>
<th>B73</th>
<th>PHG39</th>
<th>PHJ40</th>
<th>Mo17</th>
<th>Oh43</th>
<th>PH207</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Marker-based Kinship</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B14</td>
<td>SSS</td>
<td>1</td>
<td>0.08</td>
<td>0.27</td>
<td>0.35</td>
<td>0.05</td>
<td>0</td>
<td>0.01</td>
<td>0.03</td>
</tr>
<tr>
<td>B37</td>
<td>SSS</td>
<td>1</td>
<td>0.12</td>
<td>0.34</td>
<td>0.19</td>
<td>0.02</td>
<td>0.01</td>
<td>0.03</td>
<td></td>
</tr>
<tr>
<td>B73</td>
<td>SSS</td>
<td>1</td>
<td>0.11</td>
<td>0.04</td>
<td>0.03</td>
<td>0.01</td>
<td>0.04</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PHG39</td>
<td>SSS</td>
<td>1</td>
<td>0.08</td>
<td>0.03</td>
<td>0.01</td>
<td>0.03</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PHJ40</td>
<td>SSS</td>
<td>1</td>
<td>0.15</td>
<td>0.03</td>
<td>0.01</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mo17</td>
<td>NSS</td>
<td>1</td>
<td>0.02</td>
<td>0.03</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Oh43</td>
<td>NSS</td>
<td></td>
<td>1</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PH207</td>
<td>NSS</td>
<td></td>
<td>1</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Non-Stiff Stalk (NSS)</strong></td>
<td></td>
<td>0.03</td>
<td>0.03</td>
<td>0.03</td>
<td>0.04</td>
<td>0.05</td>
<td>0.09</td>
<td>0.07</td>
<td>0.16</td>
</tr>
<tr>
<td><strong>Stiff Stalk (SSS)</strong></td>
<td></td>
<td>0.23</td>
<td>0.21</td>
<td>0.22</td>
<td>0.25</td>
<td>0.1</td>
<td>0.03</td>
<td>0.02</td>
<td>0.03</td>
</tr>
<tr>
<td><strong>G-based Kinship</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B14</td>
<td>SSS</td>
<td>2</td>
<td>-0.05</td>
<td>0.38</td>
<td>-0.31</td>
<td>-0.18</td>
<td>-0.30</td>
<td>0.50</td>
<td>-0.12</td>
</tr>
<tr>
<td>B37</td>
<td>SSS</td>
<td>2</td>
<td>0.05</td>
<td>-0.31</td>
<td>-0.17</td>
<td>-0.32</td>
<td>0.58</td>
<td>0.25</td>
<td></td>
</tr>
<tr>
<td>B73</td>
<td>SSS</td>
<td>2</td>
<td>-0.29</td>
<td>-0.20</td>
<td>-0.29</td>
<td>-0.04</td>
<td>-0.17</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PHG39</td>
<td>SSS</td>
<td>2</td>
<td>-0.14</td>
<td>-0.28</td>
<td>-0.23</td>
<td>0.21</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PHJ40</td>
<td>SSS</td>
<td>2</td>
<td>-0.29</td>
<td>-0.23</td>
<td>-0.10</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mo17</td>
<td>NSS</td>
<td>2</td>
<td>-0.32</td>
<td>-0.30</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Oh43</td>
<td>NSS</td>
<td>2</td>
<td>-0.03</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PH207</td>
<td>NSS</td>
<td></td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Non-Stiff Stalk (NSS)</strong></td>
<td></td>
<td>-0.20</td>
<td>-0.19</td>
<td>-0.22</td>
<td>0.07</td>
<td>0.06</td>
<td>0.17</td>
<td>-0.18</td>
<td>-0.05</td>
</tr>
<tr>
<td><strong>Stiff Stalk (SSS)</strong></td>
<td></td>
<td>0.36</td>
<td>0.33</td>
<td>0.35</td>
<td>-0.20</td>
<td>-0.15</td>
<td>-0.26</td>
<td>0.40</td>
<td>0.09</td>
</tr>
<tr>
<td><strong>Marker-based Kinship</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B14</td>
<td>SSS</td>
<td>1</td>
<td>0.06</td>
<td>0.22</td>
<td>0.04</td>
<td>0.13</td>
<td>0.03</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>B37</td>
<td>SSS</td>
<td>1</td>
<td>0.22</td>
<td>0.04</td>
<td>0.13</td>
<td>0.03</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>B73</td>
<td>SSS</td>
<td>1</td>
<td>0.03</td>
<td>0.06</td>
<td>0.02</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>PHG39</td>
<td>SSS</td>
<td>1</td>
<td>0.01</td>
<td>0.02</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>PHJ40</td>
<td>SSS</td>
<td>1</td>
<td>0.01</td>
<td>0.03</td>
<td>0.02</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mo17</td>
<td>NSS</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Oh43</td>
<td>NSS</td>
<td></td>
<td>1</td>
<td>0.04</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PH207</td>
<td>NSS</td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Non-Stiff Stalk (NSS)</strong></td>
<td></td>
<td>0.00</td>
<td>0.01</td>
<td>0.00</td>
<td>0.00</td>
<td>0.01</td>
<td>0.07</td>
<td>0.05</td>
<td>0.16</td>
</tr>
<tr>
<td><strong>Stiff Stalk (SSS)</strong></td>
<td></td>
<td>0.25</td>
<td>0.31</td>
<td>0.34</td>
<td>0.24</td>
<td>0.11</td>
<td>0.02</td>
<td>0.00</td>
<td>0.00</td>
</tr>
</tbody>
</table>

NSS: non-stiff stalk heterotic group; SSS: stiff stalk heterotic group
Figure 3.1. Genetic diversity within and between the stiff stalk (SSS) and non-stiff stalk (NSS) heterotic groups. Diversity was estimated as 1 minus the kinship coefficient calculated with A) a marker-based and B) a pedigree-based distance matrix. Horizontal black line in the box and red diamond corresponds to the median and mean diversity per group, respectively.
Figure 3.2. Dendrogram based on UPGMA clustering from the marker-based genetic distance matrix computed from the posterior probability of IBD status at each locus. This circular dendrogram highlights the heterotic subgroup structure found for the 52 inbred lines. Major heterotic subgroups are labeled broadly based on their relevant ancestor line.
Figure 3.3. Dendrogram based on UPGMA clustering from a kinship matrix computed with the VanRaden’s (2008) methodology. This circular dendrogram highlights the heterotic subgroup structure found for the 52 inbred lines. Major heterotic subgroups are labeled broadly based on their relevant ancestor line.
Figure 3.4. Dendrogram based on UPGMA clustering from the pedigree-based genetic distance matrix. This circular dendrogram highlights the heterotic subgroup structure found for the 52 inbred lines. Major heterotic subgroups are labeled broadly based on their relevant ancestor line.
Figure 3.5. Proportion of the genome classified as IBD between ex-PVP lines and each of the key ancestor lines (B14, B37, B73, PHJ40, PHG39, PHK29, LH132, Mo17, Oh43, and PH207) for each heterotic subgroup identified with marker-based dissimilarity values. Non-IBD category (grey bars) correspond to SNPs not included within an IBD segment between the line and key ancestor.
Figure 3.6. Histogram showing the density distribution in logarithmic scale (\(\log_{10}[\text{Mb}]\)) of the identity by descent segments detected among the key ancestor lines and ex-PVP lines within A) stiff stalk (SSS) and B) non-stiff stalk (NSS) heterotic group. The red dashed lines represent the quantiles 25% (Q1) and 75% (Q3). The grey area represents the density probability for IBD segment lengths. Inside box is the summary statistic with Q2 representing the median.

Figure 3.7. Distribution of the number of generations of recombination an IBD segment was originated for the stiff stalk (SSS) and non-stiff stalk (NSS) heterotic group. The curly bracket corresponds to the cumulative percentage for inbreeding created between 35 and 135 generations.
References


Hill W, Robertson A (1968) Linkage disequilibrium in finite populations. Theoretical and applied genetics 38:226-231


Laurie-Ahlberg CC, Weir BS (1979) Allozymic Variation and Linkage Disequilibrium in Some Laboratory Populations of DROSOPHILA MELANOGASTER. Genetics 92:1295


Malécot G (1948) Mathématiques de l’hérédité


Rudan I, Campbell H (2004) Five reasons why inbreeding may have considerable effect on post-reproductive human health. Collegium antropologicum 28:943-950


Wang J (2016) Pedigrees or markers: Which are better in estimating relatedness and inbreeding coefficient? Theoretical population biology 107:4-13


Wright S (1921) Systems of mating. II. The effects of inbreeding on the genetic composition of a population. Genetics 6:124-143


CHAPTER 4. PRINCIPLES BEHIND COMMON ALGORITHMS IN GENOME-WIDE ASSOCIATION STUDIES (GWAS)

Fernando S Aguilar¹, Alejandro Ledesma¹, Thomas, Lubberstedt¹, and Jode Edwards²

¹ Department of Agronomy, Iowa State University, Ames, IA 50011, USA
² USDA-ARS Corn Insects and Crop Genetics Research Unit, Ames, IA 50011, USA

Modified from a manuscript to be submitted to Theoretical and Applied Genetics

Abstract

Genome-wide association studies (GWAS) are aimed for the identification of associations between genetic variants and phenotype. Several methods have been implemented, with the mixed linear models (MLM) being amply accepted and used. The goal of this review was to introduce the basic statistical concepts behind some common methodologies used for normally and non-normally distributed traits and to test them using a real data example with 487 double haploid lines for plant height. The real data example showed that SUPER methodology was influenced by several moderately significant SNPs, with the QQ plot having an early deviation from the expected -log₁₀[P values]. FarmCPU model identified 9 significant SNPs, however, only the 5 more significant SNPs were associated with plant height because the other 4 SNPs were correlated or influenced by these SNPs resulting in false positive associations. Using FarmCPU three genes Zm00001d008266, Zm00001d008266, and Zm00001d005951 located on chromosome 7, 8, and 2, respectively, were associated with plant height. Gene Zm00001d008266 was involved in brassinosteroids signaling, Zm00001d008266, in auxin catabolism, and Zm00001d005951 was associated with growth regulators in the response to water deficit. Validation of these genes will be required for further use in breeding programs.
Introduction

With the advent of molecular markers to identify genetic variants, an increased interest in testing associations between markers and a variety of traits was observed. These analyses were initially referred to as “association studies” and were primarily used in case-control designs where the frequency of a genotype at a marker was used to find an association with the trait (Hirschhorn et al. 2002). Subsequent analysis targeting associations across the whole genome gave way to “Genome-Wide Association Studies” (GWAS) as a specific type of association analysis (Hirschhorn et al. 2002; Liu 2014). The advantage of GWAS over conventional association studies relies on the lack of dependence on the genomic location of the genetic variant, allowing the identification of associations even in the absence of evidence regarding the function of causal genes (Hirschhorn and Daly 2005). However, the success of GWAS depends on the ability to screen hundreds or thousands of individuals for a large number of polymorphic markers, such that the probability of finding markers associated or highly correlated (i.e. in linkage disequilibrium) with the causal allele increases (Hirschhorn and Daly 2005).

Improvements in genotyping technologies have facilitated the rapid discovery of common polymorphisms. With this development of technologies, genotyping costs per marker decreased making GWAS studies more feasible. Nevertheless, GWAS faced new challenges because of an increasing number of markers \( (p) \) compared to the number of observations (i.e. large \( p \) small \( n \)), requiring the incorporation of statistical methodologies to explore relationships between markers and phenotype without increasing the rate of false-positive associations. Several methods have been implemented during the time such as penalized likelihood, variable selection, genomic control (GC) (Devlin et al. 2001), structured associations (SA) (Falush et al. 2003; Pritchard et al. 2000a), principal component analysis (PCA) (Price et al. 2006), multidimensional scaling (MDS), and the use of mixed linear models (Li and Zhu 2013; Mei and Wang 2016; Yu et al.
Among them, the mixed linear models (MLM) has been amply used and accepted, because its ability to incorporate the polygenic relationship between individuals as random effects while including population structure, marker, and non-marker effects as fixed terms in the model (Aulchenko et al. 2007; Li and Zhu 2013; Zhang et al. 2010a). Using Henderson (1984) notation, the mixed linear model could be expressed as:

\[ y = X\beta + Zu + e \]

Where \( y \) represents a vector of the response variable for each pedigree with a dimension of \( n \) by 1, \( \beta \) the collection of markers and non-markers fixed effects, \( u \) the random additive polygenic effect, \( X \) and \( Z \) the design matrices for the fixed and random terms respectively, and \( e \) the random residuals (Aulchenko et al. 2007; Henderson 1984; Zhang et al. 2010a). Under this model, both \( u \) and \( e \) are assumed to be independent and normally distributed (Aulchenko et al. 2007). The variance-covariance for random polygenic effects is defined as \( K\sigma_u^2 \), where \( K \) describes the degree of resemblance between two individuals or kinship and \( \sigma_u^2 \) the additive genetic variance due to polygenes (Aulchenko et al. 2007; Yu et al. 2006; Zhang et al. 2010a). The residual variance is defined as \( \sigma_e^2 I \), where \( \sigma_e^2 \) is the residual variance and \( I \) the identity matrix (Aulchenko et al. 2007; Henderson 1984), suggesting a common residual variance across all covariates or homoscedasticity.

Yu et al. (2006) proposed a model with the inclusion of both population structure (\( Q \)) and kinship matrix (\( K \)) in the well-known \( QK \)-model, which increases the goodness of fit compared to other models. Expanding Henderson (1984) notation to include population structure in the model and splitting the fixed effects into the ones associated with markers and non-markers, the \( QK \)-model can be written as:

\[ y = X\beta + S\alpha + Qv + Zu + e \]
Where $\beta$ now refers to all non-marker fixed effects, $\alpha$ to the marker fixed effects, $v$ to the population of origin fixed effect, $u$ to the polygenic random effect, and $e$ to the random residuals (Yu et al. 2006). Likewise, $X$ and $S$ are the incidence matrices for the non-marker and marker fixed effects, $Q$ a matrix that relates the population of origin effect ($v$) with the response variable $y$, and $Z$ the incidence matrix for the random polygenic effects (Yu et al. 2006). The efficiency in finding associations increases significantly with the use of the $QK$-model, however, the time to solve the normal equations increases drastically (Aulchenko et al. 2007). According to Zhang et al. (2010a), solving the normal equations with large data sets is computationally intensive, increasing in time as the number of individuals, the number of markers, and the number of iterations increase. Hence, direct implementation of a mixed linear model to account for associations becomes impossible when the number of individuals to be used ranges in the thousands, requiring hours or days up to several years before finding some associations (Zhang et al. 2010a). Several approaches were proposed to develop a $QK$-model that relies on grouping the individuals based on identity-by-descent, fitting a two-step approach to solving the normal equations, and/or using spectral decomposition to reduce dimensionality (Aulchenko et al. 2007; Kang et al. 2008a; Lippert et al. 2011; Liu et al. 2016; Segura et al. 2012; Zhang et al. 2010a). Therefore, the goal of this article was to introduce the basic statistical concepts behind some common methodologies used for normally and non-normally distributed traits. The review is organized into seven sections. Section 1) correspond to the introduction, 2) to a description of population structure, 3) to a description of the matrix of relationship, 4) to the principles behind some methods used to find associations, 5) to a real data example, and 7) to a summary.
Population structure

When some individuals within a population are more related by ancestry than it is expected under panmictia, the population is said to have a structure (Astle and Balding 2009; Hirschhorn and Daly 2005). Population structure (PS) was often interpreted by phenotypic or geographic stratification with a model assuming the distribution of individuals into separate subpopulations (Astle and Balding 2009). Within a subpopulation, assortative mating could lead to allelic differentiation between individuals producing a within-subpopulation structure, containing past information about admixture, migration, immigration, genetic drift, random drift, causal mutations, and/or selection among others (Price et al. 2006). The classification of individuals into genetically similar groups can be done using principal component analysis (PCA), distance-based methods, or model-based methods (Price et al. 2006; Pritchard et al. 2000a). PCA relies on the reduction of dimensionality using principal components, trying to maximize the genetic variability (Price et al. 2006). Under this analysis, each principal component will account for a percentage of the total genetic variance by grouping the individuals into clusters with similar genetic information. After reducing dimensionality, a linear regression model is fitted to each of the axes of variation, and the residuals are extracted to compute associations (Price et al. 2006). One of the advantages of PCA is the avoidance of any prior information about individual ancestries, population of origin, and/or assumptions about the data, being able to handle genome-wide data for thousands of individuals (Paschou et al. 2007).

Distance-based methods rely on the computation of an $n \times n$ distance matrix with diagonal elements expressing no distance and off-diagonal elements providing information about the genetic distance between two individuals (Pritchard et al. 2000a). This type of analysis does not allow, in an easy way, the incorporation of any additional information of the individuals, as well as it is highly dependent on the distance measurement algorithm and graphical representation.
that is chosen (Pritchard et al. 2000a). Hence, distance-based methods are used mainly as descriptive analysis and are not considered an accurate tool to quantify the structure of a population. Alternatively, model-based methods rely on the assumption that draws of individuals from a cluster (i.e. subpopulation) corresponds to a random draw from a parametric model (Pritchard et al. 2000a). Model-based methods have the advantage of allowing modeling individuals based on no admixture or admixture between individuals, on linkage between loci, and/or on the use of prior information from the population (Pritchard et al. 2000a).

Assessing population structure using model-based methods relies on a Bayesian approach to identify populations from the data and assign individuals to a population by applying a Markov Chain Monte Carlo (MCMC) algorithm (Porras-Hurtado et al. 2013; Pritchard et al. 2000a). The algorithm starts by creating \( t \) groups with \( n \) individuals and \( P \) allele frequencies at each locus, followed by a re-assignation of individuals based on those \( P \) frequencies (Porras-Hurtado et al. 2013). The process is iterated several times by setting a user-defined burn-in and MCMC chain (Porras-Hurtado et al. 2013; Pritchard et al. 2000a). A model-based method without admixture can be implemented by assuming no knowledge of the populations under study, meaning that the probability that an individual belongs to the population is the same for all \( t \) populations (Pritchard et al. 2000a). Further, this model assumes that each individual comes from a different population and that no crossing between them was present to create that genotype (Pritchard et al. 2000a). Another option is to implement a model that allows admixture between populations, suggesting that some individuals have a common ancestor from multiple populations (Porras-Hurtado et al. 2013). For both models, the allele frequencies in all populations (\( P \)) follow a Dirichlet distribution (\( Dir(\alpha) \)) whose entries move between zero and one (Pritchard et al. 2000a). One particularity of this distribution is that when large values of \( \alpha \)
are used (~ 1), the model will assume that the alleles at each locus come from all $t$ populations (admixture), while for small values of $\alpha$ (~ 0) the alleles will come from a single and different population (Pritchard et al. 2000a). Model-based methods assume Hardy-Weinberg equilibrium within populations and linkage equilibrium between loci (Pritchard et al. 2000a). For plant breeding populations, the assumption of linkage equilibrium can sometimes be violated, forcing the implementation of a model that allows linkage disequilibrium (LD) to account for correlations by ancestry (Falush et al. 2003; Porras-Hurtado et al. 2013). Falush et al. (2003) propose a methodology that accounts for two types of linkage disequilibrium: mixture and admixture LD. The first one is caused by the variation in ancestry, which leads to correlations between markers, while the second was because of the ancestry along the chromosome, resulting in an additional linkage between linked markers (Falush et al. 2003).

False associations, that is, associations between markers not linked to any causative loci, can be caused by the variation in allele frequencies across a subpopulation (Pritchard et al. 2000b). High frequencies for a particular allele, resulting from an increase in the number of related individuals in a subpopulation, could lead to false associations (Price et al. 2006; Pritchard et al. 2000b). Therefore, accounting for population structure will reduce the probability of false associations by reducing the connection between rare alleles or non-casual SNPs and the phenotype. Several methods to estimate population structure have been proposed during the time but PCA and model-based methods, commonly calculated with STRUCTURE software, have been amply accepted (Price et al. 2006; Pritchard et al. 2000a). However, as pointed out by Lipka et al. (2015), both PCA and STRUCTURE provide, on average, the same results, with the latter requiring higher computational resources because of the number of iterations needed for the Bayesian clustering approach (Lipka et al. 2015).
Matrix of relationship

Kinship quantifies the degree of relationship between two individuals (Falconer and Mackay 1996). This relationship measures the probability that at a single locus, two alleles sampled at random from two individuals, one from each, are identical copies coming from a common ancestor (Bernardo 2010; Bernardo et al. 1996; Falconer and Mackay 1996; Malécot 1948; Wright 1922). Kinship values were originally estimated from pedigree information, however, with the advent of genotyping techniques, new procedures to compute the relatedness between individuals have been proposed. Kinship matrices measure the probability that two individuals share alleles that are identical by descent (Backus and Gilpin 2002; Bernardo 2010; Falconer and Mackay 1996). Pedigree-based kinship estimates provide the probability conditioned on pedigree information, while marker-based estimators calculate the relationship based on genotypic information. Marker-based kinship coefficients rely on the marker allele frequencies in the reference population, which in practice is not known (Wang 2014). Therefore, these estimates commonly use the sample of genotyped individuals as the reference population, resulting in the estimation of the correlation that two homologous genes within or between individuals are shared by descent (Wang 2014).

Marker-based kinship coefficients opposed to pedigree-based coefficients can result in negative values. Wang (2014), states that negative values for the kinship coefficient could be interpreted as a lower probability that two homologous alleles are shared by descent compared to the probability obtained when the two alleles are taken at random from the reference population. The use of molecular markers in the computation of kinship has the advantage of identifying the number of identical markers shared between individuals, however, these estimates could confound alleles identical by state (IBS), that is alleles that are physically the same allele, and alleles that are identical by descent (IBD). In that regard, Bernardo (1993) proposed a similarity
index “S_{AB}” followed by a correction for the proportion of IBS alleles. He indicates that the proportion of shared loci between a pair of individuals (S_{AB}) is a function of the true kinship (f_{AB}') and of alleles that are IBS but not IBD (\delta_{AB}). In practice values for \delta_{AB} between two individuals are unknown and thus \delta_{AB} is calculated by measuring the average proportion of alleles shared between an individual and the two parental lines (\delta_{AB} = 0.5(\delta_A + \delta_B)). Letting X be an n by m matrix containing the SNP markers of n individuals at m loci, the relationship matrix (K) can be calculated as X^TX (Endelman 2011; Endelman and Jannink 2012; VanRaden 2008).

As a result, diagonal elements in the K matrix will measure the number of homozygous loci for each individual and off-diagonal elements the degree of similarity between individuals (Endelman 2011; Endelman and Jannink 2012; VanRaden 2008). VanRaden (2008) proposed the computation of an additional matrix Z. In their computation, Z was the result of the difference between the marker matrix X and a matrix in which all its elements corresponds to 2(p_i - 0.5), where p_i was the frequency of the second allele at marker i (VanRaden 2008). Therefore, VanRaden (2008), compute the kinship matrix as K = \frac{ZZ'}{2\Sigma p_i(1-p_i)}, in which the denominator was used to scale the matrix to make it analogous to the pedigree-based kinship matrix (VanRaden 2008). Another approach was proposed by Endelman and Jannink (2012), in which an algorithm to compute the K matrix was developed by setting the mean of the diagonal elements to (1 + f), where f was the coefficient of inbreeding of the actual population. Thus, Endelman and Jannink’s realized kinship matrix was computed as \bar{K} = \frac{XX'}{2\Sigma_k p_k q_k}, where p_k and q_k represents the allele frequencies at locus k (Endelman and Jannink 2012). In their estimation the diagonal elements of the \bar{K} should range between 1, for non-inbred individuals, and 2, for inbred lines, however, they found that for some cases the values were bigger than 2 (Endelman and Jannink 2012).
Marker-based kinship estimates correspond to an optimal estimator of the genome-wide variance-covariance matrix as long as the number of markers exceeds the number of genotypes (Endelman and Jannink 2012). Thus, once the number of markers decreases, a shrinkage to the estimate based on the weighted average is produced, with shrinkage factor $\Sigma = \vartheta T + (1 - \vartheta)C$, where $\vartheta$ corresponds to the shrinkage intensity that ranges between 0 and 1, $C$ to the sample covariance, and $T$ to a low-dimensional model with fewer parameters than $\Sigma$ (Endelman and Jannink 2012). After including shrinkage estimates, Endelman and Jannink (2012) defined a new $K$ matrix as $K^* = \frac{\delta(S_{ij}) I + (1 - \delta) S^* + XX'}{2p_h q_k}$. Based on this estimation, the kinship corresponded to a matrix of relationships between individuals based on the probability of IBS at casual loci instead of IBD (Yu et al. 2006).

As stated in the literature, alleles that are identical by descent are, by definition, identical by state, but not all IBS alleles are copies coming from a common ancestor. Thus, an overestimated value of IBD between a pair of individuals could be produced when using IBS-based methodologies. Two major methods have been proposed to account for IBD using genotypic data. The first corresponds to an observational genotype or haplotype counting algorithm that looks for matches along chromosomes by moving a user-defined sliding window (Browning and Browning 2012; Ceballos et al. 2018). This approach identifies an IBD segment based on the number of consecutive homozygote SNPs and/or the length of the segment (Browning and Browning 2012; Ceballos et al. 2018). With the identified IBD segments, a coefficient of inbreeding and kinship is computed by dividing the length of IBD segments across all chromosomes by the total length of the genome covered with the SNPs (McQuillan et al. 2008). The second methodology corresponds to a hidden Markov model to estimate the IBD status of each SNP (Browning and Browning 2012; Ceballos et al. 2018). This type of model
works as a normal Markov chain in which the internal states are not known and can only be estimated using a prior, emission, and transitions probabilities. Several programs using either the genotype counting algorithm like GERMLINE (Gusev et al. 2008), PLINK (Purcell et al. 2007), and BEAGLE (Browning et al. 2018), or on an HMM model like BCFtools/ROH (Narasimhan et al. 2016), hmmIBD (Schaffner et al. 2018), ZooROH (Druet and Gautier 2017), ChromIBD (Druet and Farnir 2011), and IBDLD (Han and Abney 2011, 2013), among others, have been proposed to increase the power to detect IBD segments without increasing the computational burden. All these programs can produce a realized kinship matrix based on the number of SNPs that are classified as being identical by descent between individuals. For instance, IBDLD program version 3.38 (Han and Abney 2011, 2013), uses a probabilistic approach to estimate the posterior probability for the 9 Jacquard’s condensed coefficients of identity-by-descent (Jacquard 1974) at each SNP. This program then computes an empirical genome-wide kinship matrix by a weighted average of the form $m^{-1} \sum_{i}^{m} (\Delta_{1,i} + \frac{1}{2} (\Delta_{3,i} + \Delta_{5,i} + \Delta_{7,i}) + \frac{1}{4} \Delta_{8,i})$, where $\Delta_{r,i}$ is the posterior probability for the condensed identity coefficient $r$ ($r = 1,3,5,7,8$) at locus $i$, and $m$ to the total number of markers (Han and Abney 2013; Jacquard 1974). Studies have been conducted to estimate the advantage of using a realized IBD kinship matrix compared to the pedigree-based or marker-based kinship matrices in association studies, concluding that there is a high improvement by using realized IBD kinships when pedigree information of the subjects is deficient (Browning and Browning 2012; Gauvin et al. 2014; Peralta et al. 2018; Speed and Balding 2015).
Genome-wide association studies (GWAS) methodologies

Compressed mixed linear model (CMLM) and population parameters previously determined (P3D).

Zhang et al. (2010a) proposed a methodology based on a mixed linear model in which the random effects are compressed into genetically similar individuals. CMLM approach, which refers to the hierarchical grouping of individuals into clusters, reduces the dimensionality of the kinship matrix by clustering the total number of individuals \(n\) into \(s\) groups (where \(s \leq n\)) according to the relatedness among them (Zhang et al. 2010a). Therefore, the dimensionality of \(u\) is reduced from \(n\) to \(c = n/s\) individuals, and the kinship matrix \((K)\) is replaced by the relationship between groups (Zhang et al. 2010a). The CMLM method follows a model of the form (Zhang et al. 2010a):

\[
y = 1\mu + Q\nu + X\beta + Zu^* + e
\]

Where \(u^*\) corresponds to the vector of random polygenic effects between groups or clusters (Zhang et al. 2010a), while all other terms correspond to the ones described earlier (section 1). CMLM assumes that the random effects \(u^*\) follows a normal distribution with mean zero and variance-covariance \(K\sigma_u^2\), where \(K\) is defined as the average kinship among the individuals from a cluster (Zhang et al. 2010a). By using this approach, the time of processing the MLM can be reduced by a factor \(c^3\) giving a final estimated time proportional to \(mpc^3\), for \(m\) representing the number of markers and \(p\) the iterations (Zhang et al. 2010a). The hierarchical clustering of individuals can be done by using the unweighted pair-group method with arithmetic averages (UPGMA), the unweighted pair-group centroid (UPGMC), the complete linkage (COM), the Lance-Williams flexible beta method (FLE), the McQuintty’s similarity, the weighted pair-group method with arithmetic averages (WPGMA), the weighted pair-group method with centroids (WPGMC), the single linkage (SIN), and Ward’s method (WAR), among
others (Li et al. 2014; Zhang et al. 2010a). CMLM algorithm commonly uses the UPGMA method for clustering. UPGMA starts by creating a correlation matrix between the genotypes, followed by the selection of the two individuals with the highest correlation coefficient to create the first two clusters (Sokal 1958). New individuals are included in each cluster by considering the average correlation with the members of each group (Sokal 1958). Individuals will be grouped thenceforth into the already existing clusters until the average correlation between the individual and the groups is less than the average correlation within the groups (Sokal 1958). Finally, new groups of individuals will be created hierarchically by measuring the average correlation between groups and within groups (Sokal 1958).

As an addition to the CMLM algorithm, Zhang et al. (2010a), developed the Population Parameters Previously Determined (P3D) algorithm to estimate the variance-covariance parameters without iterations. The P3D algorithm solves an MLM in two steps. The first consists of fitting a CMLM algorithm without including the marker effects, while the second step uses an approach like a Bayesian analysis with prior information assumed to be the population parameters estimated in the first step (Zhang et al. 2010a). The algorithm provides estimators of the variance components ($\hat{\sigma}_u^2$ and $\hat{\sigma}_e^2$) and the cluster factor ($\hat{c}$) in the first step by maximizing the joint likelihood function $L(y|v, u, \sigma_u^2, \sigma_e^2, c)$ (Zhang et al. 2010a). In the second step, the algorithm uses this estimator as fixed population parameters to fit an MLM (Zhang et al. 2010a). P3D algorithm maximizes, in the second step the likelihood $L(y|\beta, v, u, \hat{\sigma}_u^2, \hat{\sigma}_e^2, \hat{c})$ to find estimates for the fixed marker effect (Zhang et al. 2010a). By holding the variance components and clustering effects as fixed constants in the CMLM, the iteration process to compute them is avoided and thus, the process is accelerated (Zhang et al. 2010a). Higher statistical powers have been reported when using compression, however, once an optimum is reached, the statistical
power of the CMLM-P3D method reduces to the level of a general linear model (GLM) (Zhang et al. 2010a). According to Zhang et al. (2010a), better control of false-positive associations was found when the compression level \((c)\) was between 1.5 and 10 groups (Zhang et al. 2010a). However, Kang et al. (2015), states that the CMLM-P3D method provides generally different SNP ranks and higher SNP P-values \((-\log_{10}[\text{P-value}])\) than an MLM method, suggesting that extra caution should be used when small population sizes are used. Nonetheless, when choosing a compression level that produces a statistical power comparable with that of the standard MLM, a reduction in time up to 103-fold is achieved (Zhang et al. 2010a).

**Efficient mixed model association (EMMA)**

Efficient Mixed Model association algorithm allows the estimation of constants (marker effects) and variance components (residual and polygenic effects) when the data is to be explained by the use of both fixed and random effects (Patterson and Thompson 1971). This approach makes use of spectral decomposition to avoid matrix operations in each iteration and to reduce the time from \(n^3\) to a linear function of \(n\) when solving a QK-model (Kang et al. 2008a). EMMA enables the test of fixed effects and variance components by maximizing the joint likelihood of all possible contrast of the variance components (Patterson and Thompson 1971). The maximization is done only on the variance components because the information from fixed effects provides information for estimating the random and residual variance (Patterson and Thompson 1971). Thus, a full \((l_F)\) and a restricted \((l_R)\) log-likelihood function (restricted because of the lack of fixed effects), are formulated as (Kang et al. 2008a; Patterson and Thompson 1971):

\[
l_F(y; \beta; \sigma_u; \sigma_e^2 / \sigma_u^2) = \frac{1}{2} \left( -n \ln(2\pi\sigma_u^2) - \ln(H) - \frac{1}{\sigma_u^2}(y - X\beta)'H^{-1}(y - X\beta) \right)
\]
\[ l_R(y; \sigma_u^2; \sigma_e^2 / \sigma_u^2) = l_F(y; \hat{\beta}; \sigma_u^2; \sigma_e^2) + \frac{1}{2} (q \ln(2\pi \sigma_u^2) + \ln|X'X| - \ln|X'H^{-1}X|) \]

Where \( H \) corresponds to \( ZKZ' + \frac{\sigma_e^2}{\sigma_u^2} I \) and \( q \) to the degrees of freedom associated with the fixed effects (Kang et al. 2008a; Patterson and Thompson 1971). The optimal variance components for both full \( (\sigma_F^2) \) and restricted \( (\sigma_R^2) \) likelihood functions are obtained as:

\[ \sigma_F^2 = \frac{(y-X\hat{\beta})'H^{-1}(y-X\hat{\beta})}{n} \]
\[ \sigma_R^2 = \frac{(y-X\hat{\beta})'H^{-1}(y-X\hat{\beta})}{n-q} \]

respectively (Kang et al. 2008a). In this approach, estimates for \( \hat{\sigma}_u^2 \) and \( \hat{\sigma}_e^2 \) are obtained by maximizing the restricted likelihood function \( l_R \), while estimates for \( \beta \) (\( \hat{\beta} \)) are obtained by maximizing the full likelihood function where \( \sigma_u^2 \) and \( \sigma_e^2 \) are assumed to be the ones estimated with the restricted likelihood function (Patterson and Thompson 1971). The variance-covariance matrix for the restricted likelihood can be, using spectral decomposition, equated to \( S'HS = P \left[ diag(\lambda_1 + \delta, \ldots, \lambda_{n-q} + \delta) \right]P' = PSP^T \), where \( \delta \) corresponds to \( \sigma_e^2/\sigma_u^2 \), and \( P \) to an \( n \times (n-q) \) matrix in which the columns are orthogonal vectors of \( S \) (Kang et al. 2008a; Patterson and Thompson 1971). According to Patterson and Thompson (1971), \( S \) and \( P \), are idempotent and symmetric matrices that can be further expressed as \( S = AA' \) and \( P = AB \) for all \( A \) and \( B \) corresponding to matrices of dimensions \( n \times (n-q) \). Thus, the maximum-likelihood (ML) or restricted maximum-likelihood (REML) for the variance components are obtained by optimizing the functions,

\[ f_F(\delta) = \frac{1}{2} \left[ n \ln \left( \frac{n}{2\pi} \right) - n - n \ln \left( \sum_{s=1}^{n-q} \frac{\eta_s^2}{\lambda_s + \delta} \right) - \sum_{i=1}^{n} \ln(\xi_i + \delta) \right] \]

\[ f_R(\delta) = \frac{1}{2} \left[ (n-q) \ln \left( \frac{n-q}{2\pi} \right) - (n-q) - (n-q) \ln \left( \sum_{s=1}^{n-q} \frac{\eta_s^2}{\lambda_s + \delta} \right) - \sum_{s=1}^{n-q} \ln(\lambda_s + \delta) \right] \]

Where \( (\xi_i + \delta) \) corresponds to the diagonal elements of the \( H \) matrix, \( \eta_s \) to the product of \( Ay \), and \( (\lambda_s + \delta) \) to the diagonal elements of \( S'HS \) (Kang et al. 2008a). Differentiating these
functions concerning the random and fixed effects will give estimates for $\hat{\beta}$, $\hat{\sigma}_u^2$, and $\hat{\sigma}_e^2$ which by an iterative processes will produce estimates for the marker effects (Kang et al. 2010a; Kang et al. 2008a; Patterson and Thompson 1971). Once the spectral decomposition is computed, this procedure solves the derivatives iteratively without computing again the spectral decomposition, accelerating the estimation of the variance components and fixed effects while reducing the computational cost at each iteration (Kang et al. 2008a). EMMA approach combines a grid search with the Newton-Raphson algorithm to solve the normal equations for $\delta$, optimizing the likelihood functions and reducing the occurrence of false positives (Kang et al. 2008b). EMMA algorithm significantly reduces the time to analyze a whole-genome experiment, however, it works under the assumption of only one correlated variance component, named the polygenic additive effect $u$ (Kang et al. 2008b).

**Efficient mixed model association expedited (EMMAX)**

According to Kang et al. (2010a), the implementation of the EMMA algorithm in datasets with more than 1000 individuals becomes intractable, forcing the use of new approaches in estimating variance components. EMMAX methodology work under three steps: 1) producing an estimated positive-semidefinite kinship matrix that accounts for sample structure, 2) producing estimates for the additive genetic variance ($\hat{\sigma}_u^2$) and the residual variance ($\hat{\sigma}_e^2$) with EMMA algorithm, and 3) fitting a general least squared (GLS) model to test the associations between each marker and the phenotype (Kang et al. 2010a). The methodology starts by computing a positive-semidefinite pairwise kinship matrix ($\hat{S}_N$) by using either a normalized genetic relatedness between individuals or by the use of the marker-based kinship matrix (Kang et al. 2010a). With the chosen kinship ($K$) the pairwise genetic matrix $\hat{S}_N$ is computed as:

$$\hat{S}_N = \frac{(n - 1)K}{Trace(KP)}$$
Where $P = I - \frac{1}{n}$ and 1 is a vector of ones with dimensions $1 \times n$ (Kang et al. 2010a).

Then, an MLM including the estimated pairwise genetic relatedness matrix ($\hat{S}_N$) is fitted to estimate the restricted maximum likelihood estimators (REML) for the variance components $\hat{\sigma}_u^2$ and $\hat{\sigma}_e^2$, followed by a test of hypothesis for the additive genetic variance due to polygenes ($\hat{\sigma}_u^2$) (Kang et al. 2010a). In case the analysis fails to reject the null hypothesis ($H_0: \hat{\sigma}_u^2 = 0$), least squares estimates are used to find the values for the coefficients for each SNP genotyped (Kang et al. 2010a). If the null hypothesis is rejected, the best linear unbiased predictors (BLUP) are used to fit a general least squares F-test to estimate the effect of each marker (Kang et al. 2010a).

The GLS model fitted will be of the form:

$$y_i = \beta_0 + \beta_k X_{ik} + \eta_i$$

Where $\beta_0$ corresponds to a matrix that includes the intercept and/or other non-marker fixed effects included in the MLM, $\beta_k$ to the marker effects, $X_{ik}$ to the genotypic matrix of each genotype, and $\eta_i$ to the error term (Kang et al. 2010a). After finding the estimates of the variances, by using the EMMA algorithm, a GLS model in which the variance components are assumed to be known (i.e. $\hat{\sigma}_u^2$ and $\hat{\sigma}_e^2$) is fitted to find the marker and non-marker effects (Kang et al. 2010a). When the trait is controlled for some large-effect SNPs, the analysis will result in overdispersion, leading to poor associations between the markers and phenotype (Kang et al. 2008a). The fact of including a correction for sample structure ($\hat{S}_N$) allows EMMAX to correct for overdispersion when there are several markers with small effect and at the same time increase the selection of true positives associations (Kang et al. 2010a). However, EMMAX assumes small effects for each marker when computing the variance components (Kang et al. 2008a). Even when this assumption is true for most of the complex traits in plant breeding, some are controlled by a few major effects QTLs, making the algorithm to produce some biased estimates.
(Kang et al. 2008a). Consequently, Kang et al. (2008a) proposed the use of known large effect SNPs as fixed effects when computing the estimates of the variance components. As an alternative, the researcher could run an over- and under-dispersed EMMAX algorithm and select as large effect SNPs those with highly significant P-values, in other words, SNPs explaining more than 1% of the phenotypic variance (Kang et al. 2008a). The second approach is to use prior knowledge about the genetic architecture of the trait and associate those loci with large effects (Kang et al. 2008a). Similarly, EMMAX is dependent on the EMMA algorithm, which works only for the variance components associated with the random residuals and the polygenic additive variance (Kang et al. 2010a; Kang et al. 2008a). Therefore, if more than two variance components are to be included in the analysis, another methodology must be implemented to produce the estimates for the variance-covariance of the random effects before fitting the GLS model.

**Factored spectrally transformed linear mixed model (FaST-LMM)**

FaST-LMM, similar to the EMMA, uses a spectral decomposition to solve for the variance components (Lippert et al. 2011). The approach is built on the assumption that the response variable follows a normal distribution with mean $X\beta$ and variance-covariance $\sigma^2_u K + \sigma^2_e I$, where $K$ refers to the kinship matrix and $I$ to the identity matrix. FaST-LMM fits a mixed linear model with log likelihood of the form:

$$l(f(y|\sigma^2_e, \sigma^2_u, \beta)) = \log \left( \frac{1}{(2\pi(\sigma^2_u K + \sigma^2_e I))^\frac{n}{2}} e^{-\frac{1}{2}(y-X\beta)'(\sigma^2_u K + \sigma^2_e I)^{-1}(y-X\beta)} \right)$$

After multiplying every term in the log-likelihood for $(\sigma^2_u)^{-1}$ and expanding:

$$l(y|\sigma^2_u, \delta, \beta) = -\frac{1}{2} \left[ n \ln(2\pi \sigma^2_u) + \ln(|K + \delta I|) + \frac{1}{\sigma^2_u} (y - X\beta)'(K + \delta I)^{-1}(y - X\beta) \right]$$
Where $\delta$ correspond to the ratio between the residual and the additive genetic variance ($\sigma_e^2/\sigma_u^2$) (Kang et al. 2008a; Lippert et al. 2011). FaST-LMM starts by using a spectral decomposition ("Spectrally (S)" part of the model) of the kinship matrix into its linearly independent eigenvector matrices $U$ and its diagonal matrix $S$ formed with the eigenvalues of $K$ (Lippert et al. 2011). One particularity of this decomposition is that the eigenvectors are composed of orthogonal eigenvalues, making $U$ to be an orthogonal matrix that when multiplied by its transpose gives origin to the identity matrix, that is, $UU' = I$ and $U^{-1} = U'$ (Abdi 2007; Searle 1971). Replacing the kinship matrix in the log-likelihood by its spectral decomposition, the equation becomes:

$$l(y|\sigma_u^2, \delta, \beta) = -\frac{1}{2} \left[ n \ln(2\pi\sigma_u^2) + \ln(|USU' + \delta UU'|) \\
+ \frac{1}{\sigma_u^2} (y - X\beta)'(USU' + \delta UU')^{-1}(y - X\beta) \right]$$

Grouping like terms (the “Fa” or factorization term) from the variance-covariance the log-likelihood could be transformed to:

$$l(y|\sigma_u^2, \delta, \beta) = -\frac{1}{2} \left[ n \ln(2\pi\sigma_u^2) + \ln(|U(S + \delta I)U'|) \\
+ \frac{1}{\sigma_u^2} (y - X\beta)'(U(S + \delta I)U')^{-1}(y - X\beta) \right]$$

Where $(S + \delta I)$ becomes the diagonal matrix of the variance-covariance matrix (Lippert et al. 2011). Applying some algebra and knowing that the determinant of an orthogonal matrix is equal to one ($|U| = |U'| = 1$), we can rewrite the log-likelihood as the sum over $n$ terms (Lippert et al. 2011):

$$l(y|\sigma_u^2, \delta, \beta) = -\frac{1}{2} \left[ n \ln(2\pi\sigma_u^2) + \sum_{i=1}^{n} \ln(|S_{ii} + \delta|) + \frac{1}{\sigma_u^2} \sum_{i=1}^{n} \frac{([U'y]_i - [U'X]_i\beta)^2}{S_{ii} + \delta} \right]$$
The likelihood is transformed (“T” term) into a normal distribution with mean $[U’X]\beta$ and variance-covariance $\sigma^2_u (S_i + \delta)$ (Lippert et al. 2011). Differentiation of the log-likelihood function concerning $\beta \sigma^2_u$, and $\delta$ is applied to find the maximum likelihood estimator (MLE) for each parameter (Lippert et al. 2011). Logistically, the algorithm first finds the MLE for $\beta$ ($\hat{\beta}$) and use this estimate to find the MLE for the random polygenic variance $\hat{\sigma}^2_u$ (Lippert et al. 2011). Finally, the $\hat{\beta}$ and $\hat{\sigma}^2_u$ estimators are plugged into the log-likelihood function such that this becomes a function of only $\delta$ which, by optimizing it with the Brent’s method, will give a numerical one-dimensional optimizer (Lippert et al. 2011). The use of a spectral decomposition of $K$ reduces the estimation of the variance components to a linear function of $n$ instead to an $n^3$ (Lippert et al. 2011). The algorithm works by computing first a realized relationship matrix (RRM) using a sample of SNPs from the whole SNP pool to extract the eigenvalues ($S$) and eigenvectors ($U$). After this process, the algorithm rotates the phenotypes ($y$), SNPs, and covariates to create uncorrelated data (Lippert et al. 2011). In the second step, the rotated data is used to maximize the log likelihood functions to estimate the fixed SNP effects ($\beta$) and random variances (Lippert et al. 2011). In doing so, the time is reduced in the first step to $n^3$, while in the second part it will be proportional to $n^2$ (Lippert et al. 2011).

**Settlement of mlm under progressively exclusive relationship (SUPER)**

SUPER includes a new kinship matrix formed from previously known quantitative trait nucleotide (QTN) (Wang et al. 2014). Consequently, an increase in the power of SUPER is achieved by using a kinship matrix that is constructed only from influential markers (Wang et al. 2014). SUPER develops an algorithm to calculate a QTN-like SNP matrix (pseudo-QTN), which is computed by first doing a preliminary GWAS or genomic prediction analysis applying methods like GLM, MLM, or CMLM to sort SNPs based on small P-values (Wang et al. 2014).
According to Wang et al. (2014), the three preliminary methods, GLM, MLM, or CMLM, do not influence the statistical power of SUPER, however, GLM provides a faster running time compared to the other two methods. After the identification of significant SNPs, the whole genome is divided into small bins with the most influential SNPs (Wang et al. 2014). Each bin will be composed of different influential SNPs and therefore, the size and the number of bins will be different (Wang et al. 2014). With the most influential bins, the algorithm computes a kinship matrix, and the identified size (b) and number of bins (s) are used to find the restricted maximum likelihood estimators for each trait (Wang et al. 2014). Therefore, only the most associated SNPs for each bin are used as the pseudo-QTN matrix to maximize the likelihood function \( l(y|\beta, \sigma_u^2, \sigma_e^2, s, b) \) (Wang et al. 2014). Subsequently, a complementary trait specific kinship matrix is calculated by removing the pseudo-QTN-like SNPs that are in linkage disequilibrium (LD) with the testing SNP pool (Wang et al. 2014). Finally, a mixed linear model (MLM), with the complementary trait-specific kinship matrix and fixed effects (marker and non-markers effects) is fitted to find the effect of each SNP on the phenotype. SUPER can be combined with previous methodologies like FaST-LMM, P3D, EMMA, EMMAX, among others, in the final stage to solve the MLM and produce associations without losing statistical power.

The use of a specific kinship matrix increases the accuracy by using only the markers that proved to be influential in the expression of the trait (Wang et al. 2014). However, in this trait-specific kinship matrix, some markers will be confounded with the tested markers (if the SNP is part of the QTNs), reducing the statistical power of the analysis (Wang et al. 2014). The dependency of the algorithm in the threshold to find LD between QTNs and tested SNP is one of the drawbacks since when this is too high \( r^2 = 100\% \) all QTN will remain in the matrix and
thus no advantage will be produced by the use of SUPER (Wang et al. 2014). If the threshold is too low ($r^2 = 0.01\%$), most of the QTNs will be removed from the analysis and the final trait-specific kinship will no longer retain information about the trait, lowering the statistical power of the analysis (Wang et al. 2014). In summary, SUPER works as follows. First, SUPER finds the most influential SNPs (those with the lower P-value) by doing a preliminary GWAS without the random effects. Second, the significant SNPs are used to select segments in the genome to create the pseudo-QTN matrix. Third, the trait-specific kinship matrix is constructed by removing from the pseudo-QTN the QTNs that are in LD with the tested SNPs. Finally, the trait-specific kinship matrix is used to solve an MLM and find associations between the phenotype and marker SNPs.

**Fixed and random model circulating probability unification (FarmCPU)**

The rationale behind the FarmCPU algorithm is to separately fit a fixed-effects model (FEM) and a random-effects model (REM) iteratively, allowing a reduction in confounding effects, overfitting, and/or false positives (Liu et al. 2016). The algorithm starts by fitting a FEM with $m$ candidate testing markers, one at a time, to construct a pseudo QTN (Liu et al. 2016). The pseudo-QTN can be selected from previous knowledge or, if no information is available, it can be assigned as an empty set (Liu et al. 2016). In the FEM, the genetic markers are fitted as fixed effects while the pseudo QTNs are fitted as covariates following:

$$y_i = M_{i1}b_1 + \cdots + M_{im}b_m + S_{ij}d_j + e_i$$

Where $y_i$ refers to the phenotype of the $i^{th}$ individual, $M_{ij}$ to $M_{im}$ the genotypes of the selected $m^{th}$ markers (candidate pseudo-QTNs), $b_1$ to $b_m$ the marker $m$ covariate (pseudo-QTN effect), $S_{ij}$ to the $i^{th}$ genotype with the $j^{th}$ SNP, $d_j$ to the $j^{th}$ SNP fixed effect, and $e_i$ to the random residuals (Liu et al. 2016). If an empty set is used for the pseudo-QTN, the model will set to zero the covariates ($M_{im} = 0$), and the model will have only the SNP fixed effect (Liu et al. 2016). The
FEM is fitted for each genetic marker at a time and when the SNP analyzed is part of the pseudo-QTN (covariate), the model will not produce a P-value associated with its effect (Liu et al. 2016). Since this process is repeated for every marker, several P-values will be obtained for each of the pseudo-QTN SNPs, however, only the most significant will be used for that SNP (Liu et al. 2016). This process, called substitution, will provide a non-inflated P-value for each SNP because of the inclusion of the covariates in the model. After analyzing all markers, only those with a P-value lower than 1% are chosen. Selected markers will be used in a SUPER methodology to estimate the complementary trait-specific kinship matrix (Liu et al. 2016). The SUPER algorithm will be applied in a random-effects model (REM) of the form,

\[ y_i = u_i + e_i \]

Where \( y_i \) represents the phenotype of individual \( i \), \( u_i \) the genetic effect of individual \( i \), and \( e_i \) to the random residuals (Liu et al. 2016). The distribution of the random effect \( u_i \) is normal with mean zero and variance \( K \sigma_u^2 \), where \( K \) is the complementary trait specific kinship matrix computed in the FEM part of the analysis (Liu et al. 2016). After fitting the REM, the algorithm uses the trait-specific kinship matrix as the pseudo QTN for the FEM part in the second iteration (Liu et al. 2016). This process is repeated until there is no change in the pseudo QTN, quantified by the correlation between a FEM pseudo-QTN and the new REM trait-specific kinship matrix (Liu et al. 2016). By fitting a fixed model for each marker independently, the algorithm has the advantage of reducing the impact of confounding kinship effect while reducing the time to find associations to a linear function of the number of markers and individuals (Liu et al. 2016). Equally, higher control of false positives is achieved by fitting influential markers as covariates in the FEM (Liu et al. 2016). The implementation of a random-effects model to estimate the kinship matrix with the help of a SUPER algorithm, reduces the effect of overfitting obtained
when a standard MLM is used (Liu et al. 2016). As mentioned for Liu et al. (2016), the
incorporation of population structure as a covariate in the fixed effects model, either in the first
or in all iterations, increase the power to detect associations and has the advantage to eliminate
false positives because of non-genetic effects associated with the structure of the population
under study.

**Multi-locus mixed linear model (MLMM)**

Previous approaches to find associations between SNPs and phenotypes have been made
under the analysis of a single-locus test combined with population structure and polygenic
relatedness between individuals. However, these types of models leave out possible interactions
between loci, resulting in approaches not appropriate for some multi-locus complex traits
(Segura et al. 2012). Models including multiple loci have increased efficiency by including loci
that are in linkage disequilibrium and not only loci that are physically linked because of their
proximity in the chromosome (Segura et al. 2012). In GWAS, the identification of possible
causative loci could be influenced by the number of predictors (ie. Large \( p \) small \( n \) problem),
affecting the inclusion of more than one predictor (SNP) at a time (Segura et al. 2012). Segura et
al. (2012) proposed a methodology that uses a forward selection of causative markers, followed
by a backward elimination in a mixed-model regression analysis of the form:

\[
y_i = \mu + X_i \beta + Z_i u + e_i
\]

Where \( y_i \) represent the phenotype of individual \( i \), \( \mu \) the common intercept, \( X_i \) the
incidence matrix for the genotype of the individual \( i \), \( \beta \) the vector of fixed marker effects, \( Z_i \) the
design matrix for the random effects, \( u \) the polygenic random effects, and \( e_i \) the random
residuals. The random effect \( u_i \) corresponds to the polygenic effect with mean zero and variance-
covariance \( K^* \sigma_u^2 \), where \( K^* \) refers to an adjusted kinship matrix. The adjusted kinship matrix is
computed by removing the SNPs that were included as fixed effects (Segura et al. 2012). The MLMM approach starts by fitting a model with only $\mu$ and $u_i$ and then, by the use of a forward selection procedure the most significant SNPs are added to the model (Segura et al. 2012). The significance of the SNP is computed by using the $F$-to-enter test to calculate the P-values for each SNP (Ramsey and Schafer 2012; Segura et al. 2012). Thus, each step estimates the variance components $\hat{\sigma}_u^2$ and $\hat{\sigma}_e^2$ which are used in an $F$-test by using a GLS approach (Segura et al. 2012). The forward selection procedure will include all SNPs, one at a time, and select those that are significant, stopping when the addition of a new SNP does not have a significant effect in explaining the phenotypic variance, in other words, when $\frac{\hat{\sigma}_u^2}{Z'K\hat{\sigma}_u^2Z+\hat{\sigma}_e^2} \approx 0$ (Segura et al. 2012).

The model can estimate genetic ($\hat{\sigma}_g^2$), random ($\hat{\sigma}_u^2$), and residual ($\hat{\sigma}_e^2$) variances at each stage and stop the inclusion of SNPs when the genetic variance due to the new SNPs is close to zero (Trefethen 1997). After finding a suitable model with the forward selection, a backward elimination method is implemented to remove those SNP that have the smallest $F$ statistic or the higher P-value (Christensen 2011; Segura et al. 2012). The stopping point for the backward elimination approach is when the least significant cofactors (SNP) have been eliminated. For both, the forward and backward steps, the variance components for the polygenic additive variance ($\sigma^2_u$) and residual variance ($\sigma^2_e$) are estimated and used to compute the $F$-statistic (Segura et al. 2012). The adjusted kinship matrix is re-estimated each step by excluding from the matrix the markers that were included as fixed effects. Consequently, the first step in the forward selection procedure has a kinship matrix that includes all SNP markers ($K^* = K$), while the last step of backward elimination has an adjusted kinship matrix ($K^*$) in which elimination of the SNPs that were fitted as fixed effects was done, leaving only the non-significant markers to be used for the calculation (Segura et al. 2012).
**Genome-wide association studies (GWAS) for non-normally distributed traits**

Mixed linear models have proved to be an efficient way to find associations between causative loci and the underlying phenotype by controlling for cryptic relatedness and population stratification (Lipka et al. 2015). However, such a model assumes the residuals to be uncorrelated and normally distributed with mean zero and common variance-covariance $\sigma^2_e$ (homoscedastic variance). Although such assumptions are critical for GWAS analysis, they are not met for all agronomical traits, with some of them (e.g. disease status, number of aborted kernels per plant, and stalk lodging, among others) following a binary, Bernoulli, binomial, exponential, and/or gamma distribution among others. Therefore, implementing a mixed linear model for these traits will yield incorrect associations because of failure when controlling for false-positive associations (Chen et al. 2016). As an alternative, a logistic mixed regression model has been proposed as a methodology that accounts for both population structure and cryptic relatedness when the residuals do not follow a normal distribution and/or when the residual variance varies according to a covariate (i.e. heteroscedasticity) (Chen 2019; Chen et al. 2019; Chen et al. 2016; Shenstone et al. 2018; Wang et al. 2016). The logistic mixed regression model can be specified as:

$$g(p_i) = X_i b + G_i m + u,$$

Where $g(.)$ represent any link function (commonly in the exponential family), $p_i$ the probability of success for individual $i$, $X$ the incidence matrix for the fixed covariates, $b_i$ the set of non-marker fixed-effect including intercept, $G$ the matrix of genotypes (coded as 0, 1, or 2), $m_i$ the marker fixed effect, and $u_i$ the random polygenic effects (Chen et al. 2019; Milet and Perdry 2020; Stanhope and Abney 2012). Commonly the random effects are assumed to follow a multivariate normal distribution with mean zero and variance-covariance equal to $K \sigma^2_u$, where $K$
represents the kinship matrix and $\sigma_u^2$ the polygenic additive variance (Chen et al. 2019; Chen et al. 2016; Stanhope and Abney 2012). Because of the inclusion of the normal random effects ($u_i$), the likelihood function no longer follows a closed-form, forcing the use of numerical or approximated approaches like Maximum Likelihood (ML), Penalized Quasi-Likelihood (PQL), Gauss-Hermite Quadrature, Adaptive Gauss-Hermite Quadrature, or Laplace approximation, to find the marginal likelihood function (Christensen 2006; Cox and Snell 1981; Kim et al. 2013; Kleinbaum 2010; Wang et al. 2011; Whittemore and Halpern 2003). The logistic mixed regression model, using one of the numerical or approximated methods, can be implemented in SAS software (SAS Institute, Cary, NC), by using the “PROC GLIMMIX” or “PROC NLMIXED” procedures (Kim et al. 2013; Milet and Perdry 2020; Shenstone et al. 2018), however, fitting such a model for large sample sizes (> 10000 individuals) and/or for a big number of markers (> 500 K SNPs) becomes unfeasible because of a high computational burden. Therefore, several methods like generalized linear mixed model association test (GMMAT), Genome-wide LOGistic mixed model/Score test (GLOGs), and the variant-set mixed model association tests (SMMAT), among others have been proposed (Chen et al. 2019; Chen et al. 2016; Stanhope and Abney 2012). These models rely on solving the link function in two steps. The first fits a null hypothesis model, that is, a model without marker SNP effects ($m_i = 0$):

$$g(p_i) = X_i b + u_i$$

GMMAT solves this null model by using a penalized quasi-likelihood approach (Chen et al. 2016), while GLOGs uses maximum likelihood (Stanhope and Abney 2012) to estimate the polygenic additive variance. As a second step, all methods use a score test to evaluate for associations checking if there is a significant marker effect ($m_i \neq 0$) (Chen et al. 2019; Chen et al. 2016; Stanhope and Abney 2012). The use of a generalized linear mixed effects model to find
associations between causative markers and the underlying phenotype for non-normally distributed traits, has increase the power and accuracy in GWAS allowing at the same time the inclusion of population structure and cryptic relatedness.

**Real example**

To illustrate a genome-wide association studies analysis, we use a panel of doubled haploids (DH) lines representing three populations in the recurrent selection program of the Iowa Stiff Stalk Synthetic [BSSS(R)] maize population. The three populations corresponded to the unselected base population (C0), to the most advanced cycle of selection (cycle 17 or C17), and a population created by intermating individuals from C0 and C17 (C0C17). Randomly selected individuals within each population were crossed to an inducer to generate the haploid seed. Haploid seeds were germinated in plug trays and seedling underwent artificial chromosome doubling using colchicine treatment as implemented in the Iowa State Doubled Haploid facility. In total, 135, 194, and 187 DH lines were obtained for C0, C17, and C0C17 populations, respectively. The 516 DH lines plus sixteen progenitors of the BSSS synthetic population (CI187-2, AH83, Hy, 461, LE 23, I224, WD 456, Tr 9-1-1-6, Os 420, Oh 3167B, B2 (Parent of F1B1), IL12E, I-159, Fe (Parent of F1B1), CI 540, and A3G-3-3-1-3), and the line B73 were planted during summer 2019 in Burkey Agronomy Farm, Johnson Farm, and Plant Introduction Station at Ames, IA. The inbred line B73 was considered as a check and was replicated 14 times within each replicate-block in each location to have a total of 546 experimental units per replicate. The B73 inbred line was randomly distributed across the three populations, while the progenitors were included in the C0 population because its phenotypic resemblance. Each location was planted under a modified split-plot design with two replications per location where the populations (i.e. C0, C17, and C0C17) were the whole plot treatment factor and the DH lines within each population were the subplot treatment factor. This design differs from a classical
split-plot because the subplot (DH lines) were not cross-classified with the whole-plot factor and because the whole plot was assigned to more than one experimental unit in the field. The subplot experimental unit consisted of a single row spaced at 0.76 m and 3.8 m long with 15 plants. The whole-plot experimental unit was a block containing 39 subplots side-by-side. Each replicate, containing 546 subplots, was divided into 3 whole plots which were then separated into 4, 5, and 5 blocks for C0, C0C17, and C17, respectively. Each block of the whole plot was randomly assigned to a range in the field. Data was collected two weeks after pollination on an individual plot basis for plant height. Plant height was measured as the height between the base of the plant and the insertion of the flag leaf collar. Considerable variation between populations was observed, with DH lines within the C0 having the highest mean and lines within the C17 the lowest (data not shown). Data was analyzed by fitting a mixed linear model of the form:

\[ Y_{ijklmn} = \xi_i + \Gamma_{l(i)} + I_j + \omega_{ij} + \Psi_{j(k)} + \phi_{lk(j)} + \nu_{m(il)} + \delta_{n(il)} + \epsilon_{ijklmn} \]

Where \( Y_{ijklmn} \) was the response variable for range \( n \), pass \( m \), replication \( l \), DH line \( k \), population \( j \), and environment \( i \). \( \xi_i \) the effect of environment \( i \), \( \Gamma_{l(i)} \) the replication \( l \) within environment \( i \), \( I_j \) the population \( j \), \( \omega_{ij} \) the interaction between environment \( i \) and population \( j \), \( \Psi_{j(k)} \) the DH line \( k \) within the population \( j \), \( \phi_{lk(j)} \) the interaction between the DH line \( k \) within the population \( j \) and environment \( i \), \( \nu_{m(il)} \) the pass \( m \) within replication \( l \) and environment \( i \), \( \delta_{n(il)} \) the range \( n \) within replication \( l \) and environment \( i \), and \( \epsilon_{ijklmn} \) the residual error. The effects \( \Psi_{j(k)} \), \( \omega_{ij} \), \( \phi_{lk(j)} \), \( \nu_{m(il)} \), and \( \delta_{n(ij)} \) were fitted as random. Best linear unbiased predictors (BLUP) were obtained for the DH lines within population (\( \psi_{j(k)} \)) and used as a response variable for the GWAS analysis. All phenotypic analyses were done using the MIXED procedure of SAS 9.4 software (SAS Institute, Cary, NC). Genotypic information was obtained for each of the 516 DH lines using the Diversity Arrays Technology sequencing (DArT-seq) method (Jaccoud et al.)
A total of 32939 SNP markers were successfully called within the B73 reference genome version 4 (Jiao et al. 2017). The 32929 SNP markers were filtered for a minimum calling rate of 95%, minor allele frequency of 0.05, and monomorphic or multi-allelic markers. Genotypes were imputed with the linkage disequilibrium k-nearest neighbors’ imputation (LDkNNi) method of TASSEL v.5.2.52 (Bradbury et al. 2007).

Twenty-nine DH lines were discarded from the analysis because of segregation in field trials, to missing genotypic data, or levels of heterozygosity in the genotypic data above 5%. In total 132, 185, and 170 DH lines represented the C0, C17, and C0C17 populations, respectively. The 16 progenitors and the check (B73) were removed from the analysis for GWAS. The selected 487 DH lines were known to belong to the three populations (C0, C17, and C0C17), however, to exemplify, we run a principal components analysis (PCA) and a model-based clustering algorithm implemented in STRUCTURE 2.3.4 software (Pritchard et al. 2003). Principal component analysis (PCA) was performed by using the adegenet package (Jombart 2008) in R software (R Core Team 2019). Bayesian Information Criterion (BIC) (Schwarz 1978) was used to identify the optimal number of principal components by selecting the model with the lowest BIC (Figure 4.1). The model-based analysis was performed using the STRUCTURE 2.3.4 (Pritchard et al. 2003) program with K values ranging from 1 to 11. STRUCTURE was analyzed using 10000 iterations for the burn-in, 20000 for the MCMC, and 20 replications for each K. Similarly, the admixture model with correlated allele frequencies between populations and with the degree of admixture estimated from the data was used. To accelerate the estimation of K, the program structure_threader (Pina-Martins et al. 2017) was used. Structure threader was developed for the automation and parallelization of runs from Structure, FastStructure, Maverick, and ALStructure (Pina-Martins et al. 2017). The optimum number of populations (K) for the
model-based clustering algorithm was selected by using the rate of change in the likelihood function concerning K (i.e. delta K method), which breaks the slope of the distribution and shows a clear peak at the true value of K (Evanno et al. 2005). As expected, both PCA and Structure analysis showed similar results, with an optimal number of populations (or principal components) equal to 3 (Figure 4.1). The principal components, plotted in a two-dimensional plot using discriminant analysis of principal components (DAPC), showed a clear grouping of DH lines into the C0, C17, and C0C17 (Supplemental Figure 4-1). The degree of relationship among the 487 DH lines was estimated using two approaches (described in section 3). The first consisted of using a genome-wide kinship matrix computed using a probabilistic HMM framework implemented in the Identity By Descent Linkage Disequilibrium (IBDLD) program (Han and Abney 2011, 2013). This matrix was referred to as the IBDLD kinship matrix. The second approach uses genomic data to estimate a marker-based kinship coefficient computed with the AGHmatrix package (Amadeu et al. 2016). In the present study, the marker-based kinship was computed using the VanRaden (2008) methodology and was referred to as the G kinship matrix. Differences between both matrices rely on the use of estimates of IBD segments (IBDLD matrix) or SNPs identical by state (G matrix).

To determine the more suitable method for this data, we compared the GWAS models with both the IBDLD and G kinship matrix. GWAS analysis was conducted using GAPIT3 (Wang and Zhang 2018) and FaST-LMM program (Lippert et al. 2011; Widmer et al. 2014). GAPIT3 implemented the methodologies CMLM, FarmCPU, MLMM, and SUPER (Wang and Zhang 2018). The methods EMMA, P3D, and EMMAX are commonly used as algorithms to accelerate the computation time when solving the MLM, CMLM, and/or SUPER (Wang and Zhang 2018) and for that, direct analysis of these three methods was not provided in this
example. Significant SNP markers associated with each trait were identified using the Benjamini-Hochberg false discovery rate (FDR) procedure, with a level of significance of 5% (Benjamini and Hochberg 1995). Our results showed a non-significant deviation of the observed from the expected $-\log_{10}(P \text{ values})$ when the models CMLM, FaST-LMM, and MLMM with both kinship matrices were used (Figure 4.2). The models SUPER and FarmCPU, when both kinship matrices were used, showed a significant deviation of the observed from the expected values, suggesting that some SNPs were significantly associated with plant height (Figure 4.2). FarmCPU QQ plot, no matter the choice of kinship matrix, follows a straight line with a sharp upward deviated tail (Figure 4.2), suggesting the presence of significant associations (Kaler et al. 2020; Pearson and Manolio 2008). All significant SNP markers, with P values, and effect estimates are listed in table 4.1. SUPER methodology, as implemented in GAPIT3 does not provide estimates for SNP effect. For FarmCPU, the five most significant SNPs S7_25022476, S1_226374236, S8_3532960, S9_133204506, and S2_193723829 (Table 4.1), were excluded from the analysis for a second run (FarmCPU Subset) (Figure 4.2). After these SNPs were excluded, the QQ plot followed a straight line close to 1:1 between observed and expected negative log P values (Figure 4.2), suggesting that these five SNPs were truly associated with plant height and that the remaining associations were false positives correlated or influenced by these SNPs. For SUPER, the QQ plot showed an early separation of the observed from the expected $-\log_{10}(P \text{ values})$ no matter the kinship used (Figure 4.2), suggesting substantial inflation of P values with many of them being moderately significant (Table 4.1) (Ehret 2010; Kaler et al. 2020; Pearson and Manolio 2008). According to Ehret (2010) and Pearson and Manolio (2008), the early separation of the observed from the expected could be attributed to the presence of many SNPs with a moderately significant effect (rarely the case) or to the presence
of some spurious associations because of population stratification. In general, there was no difference when using the IBDLD or G matrix, but for the SUPER methodology, the former kinship leads to more significant SNPs associated with plant height (Figure 4.3). SUPER algorithm using the IBDLD matrix (SUPER + IBDLD) identified 23 SNPs while using the G matrix (SUPER + G), the algorithm identified only 3 SNPs associated with plant height (Figure 4.3). FarmCPU identified 9 significant SNPs for plant height, no matter the choice of kinship matrix (Figure 4.3). Five SNPs were found to be common between SUPER + IBDLD and FarmCPU (Figure 4.3), three of them (S7_25022476, S8_3532960, and S2_193723829) corresponding to the most significant SNPs for FarmCPU. This result goes with Kaler et al. (2020) who after comparing eight GWAS models (ANOVA, GLM, MLM, CMLM, ECMLM, SUPER, MLMM, and FarmCPU) in simulated data from maize and soybean, found that SUPER result in several false-positive associations, whereas FarmCPU efficiently controlled false positive and false negatives. Therefore, significant SNPs from the FarmCPU model were used in finding candidate genes.

The LD decayed \((r^2 = 0.2\) threshold) over a distance greater than 400 kb on average for all chromosomes (data not shown), and therefore candidate genes were assigned if they were within 800 Kb of the identified significant SNP (400 Kb to either side). The genes were annotated using the Ensembl Plants BioMart tool (Kinsella et al. 2011). The strongest association was found on chromosome 7 with the S7_25022476. This marker was located at 8.5 Kb downstream from the gene Zm00001d019268 referenced as the Lazarus 1 (LAZ1) protein gene (Kinsella et al. 2011; Portwood et al. 2019). LAZ1 is located in the tonoplast membrane, which is involved in vacuolar transport, brassinosteroids signaling, and programmed cell death signaling pathways (Liu et al. 2020; Liu et al. 2018; Malinovsky et al. 2010). Brassinosteroids
(BR) is a plant-specific hormone essential in several physiological processes like cell expansion, cell division, vegetative growth, senescence, and stress tolerance, among others (Clouse and Sasse 1998; Okleštкова et al. 2015), and thus it was not surprising to find a strong association near this gene. We found SNP S8_3532960 to be near the gene Zm00001d008266, which is involved in auxin catabolism, oxidation of toxic reductants, biosynthesis and degradation of lignin, suberization, response to environmental stresses, and oxidative stresses (Tognolli et al. 2002). According to Gallavotti (2013) auxins are one of the major plant regulators involved in plant growth and plant architecture, and for that, the gene Zm00001d008266 could be selected as a candidate gene associated with plant height. Candidate genes Zm00001d032438, Zm00001d032435, Zm00001d032439, Zm00001d032434, Zm00001d032440 on chromosome 1 were found to be near the S1_226374236 marker. These genes are in general associated with membrane solute transport in the cell and therefore, were not directly related to plant growth. SNP marker S2_193723829 in chromosome 2 was near the gene Zm00001d005951, which is associated with growth regulators in the response to water deficit (Himmelbach et al. 2002; Söderman et al. 1999). Gene Zm00001d005951 has been referred to as the ATHB6 gene in Arabidopsis and has been related with the production of a homeodomain-leucine zipper (HD-Zip) protein the number of identical markers. ATHB6 is significantly up-regulated in seedlings under water deficit and restricts cell division and/or cell differentiation in developing organs (Himmelbach et al. 2002; Söderman et al. 1999). Finally, for the SNP S9_133204506, we did not find a candidate gene related to plant height but with chloroplast accumulation or chloroplast movement in response to environmental light conditions in Arabidopsis [gene Zm00001d047516] (Kodama et al. 2010; Luesse et al. 2006). In general, we found three candidate genes Zm00001d008266, Zm00001d008266, and Zm00001d005951 located on
chromosome 7, 8, and 2, respectively, associated with plant height. Validation of these genes will be required for further use in breeding programs, requiring the production of near isogenic lines or new DH lines for expression analysis. Finally, once the genes are validated, they could be used to select donor lines with favorable allele(s).

Summary

The use of the QK-model to account for associations between the genotype and phenotype has proven to be effective in reducing spurious associations caused by population structure and relatedness between individuals. However, even when the associations produced from such model are adequate, the time and computing cost to solve the normal equations are prohibitively lasting from a couple of hours for small data sets to a couple of years for data sets consisting of thousands of individuals and markers (Kang et al. 2010b; Kang et al. 2008b). Thus, several methodologies that implements a QK-model and that use maximum likelihood (ML) or restricted maximum likelihood (REML) to obtain estimates for the polygenic additive genetic variance ($\sigma^2_u$) and the residual variance ($\sigma^2_e$) have been proposed (Kang et al. 2010b; Kang et al. 2008b; Lippert et al. 2011). EMMA, EMMAX, and PD3 algorithms work by estimating the components of variance in one step and then using them as fixed constant to estimate the effect of each marker (Wang et al. 2014). EMMA reduces the dimensionality of the analysis to one dimension by optimizing the likelihood function when the variance components have been modified to be the ratio between the residual and additive polygenic variance (Wang et al. 2014). The P3D method uses an approach to reduce the dimensionality by grouping individuals based on relatedness to produce a new kinship matrix (Zhang et al. 2010b). CMLM algorithm can use P3D to reduce significantly the time of computing associations compared to a normal QK-model, being able to reduce from $n^3$ to $c^3$ (c for the number of clusters), however, for large data sets it remains to be highly intensive (Wang et al. 2014; Zhang et al. 2010b). The SUPER methodology
makes use of any of the other proposed methodologies to solve the \textit{QK}-model. In the analysis, SUPER includes a new kinship matrix formed from previously known quantitative trait nucleotide (QTN), achieving a high power because of the realized kinship matrix is constructed only with influential markers for the trait (Wang et al. 2014), however, the algorithm has resulted in the estimation of some false positives (Kaler et al. 2020).

The real data example compared five of the GWAS models for plant height. Each model was run using a kinship matrix computed from IBD segments (IBDLD) or shared identical by state SNPs (G matrix). For the trait and DH lines used, we did not find a difference between both kinship matrices. IBD-based kinship matrices rely on the estimation of the IBD status at each locus based in an HMM, while IBS-based kinships are based on the number of identical markers between/within individuals. The genotypes used in this example constitutes double haploid lines, and therefore, the number of IBD markers should be close to the number of IBS markers, resulting in similar kinship coefficients. Based on the QQ plots, the FarmCPU model was appropriate for controlling false positives and false negatives, while SUPER was affected by the presence of several moderately significant SNPs. When removing the 5 most significant SNPs identified with the FarmCPU model, the QQ plot follows a straight line suggesting the remaining 4 SNPs were false-positive associations. From this study, we found three SNPs significantly associated with plant height, with one of them (S7_25022476) near the gene Zm00001d019268 which is related to the brassinosteroids signaling pathways.

\textbf{Data availability statement}

All datasets and scripts generated for the real data example in this study can be found on https://github.com/fsilvaag/GWAS-Review.git
Table 4.1. List of all significant SNP markers associated with plant height found with FarmCPU, and SUPER when using a kinship matrix computed from IBD segments (IBDLD) or IBS markers (G Matrix).

<table>
<thead>
<tr>
<th>Method</th>
<th>Kinship</th>
<th>SNP</th>
<th>Chromosome</th>
<th>Position (bp)</th>
<th>Effect</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>FarmCPU</td>
<td>Both</td>
<td>S7_25022476</td>
<td>7</td>
<td>25022476</td>
<td>3.8</td>
<td>1.07 x 10^{-10}</td>
</tr>
<tr>
<td></td>
<td></td>
<td>S1_226374236</td>
<td>1</td>
<td>226374236</td>
<td>2.5</td>
<td>9.42 x 10^{-07}</td>
</tr>
<tr>
<td></td>
<td></td>
<td>S8_3532960</td>
<td>8</td>
<td>3532960</td>
<td>-24.2</td>
<td>1.30 x 10^{-06}</td>
</tr>
<tr>
<td></td>
<td></td>
<td>S9_133204506</td>
<td>9</td>
<td>133204506</td>
<td>3.5</td>
<td>2.44 x 10^{-06}</td>
</tr>
<tr>
<td></td>
<td></td>
<td>S2_193723829</td>
<td>2</td>
<td>193723829</td>
<td>-3.2</td>
<td>4.31 x 10^{-06}</td>
</tr>
<tr>
<td></td>
<td></td>
<td>S9_102458787</td>
<td>9</td>
<td>102458787</td>
<td>-3.7</td>
<td>1.54 x 10^{-05}</td>
</tr>
<tr>
<td></td>
<td></td>
<td>S2_189709094</td>
<td>2</td>
<td>189709094</td>
<td>2.4</td>
<td>1.59 x 10^{-05}</td>
</tr>
<tr>
<td></td>
<td></td>
<td>S9_114880873</td>
<td>9</td>
<td>114880873</td>
<td>3.2</td>
<td>1.81 x 10^{-05}</td>
</tr>
<tr>
<td></td>
<td></td>
<td>S10_145565167</td>
<td>10</td>
<td>145565167</td>
<td>-7.2</td>
<td>2.02 x 10^{-05}</td>
</tr>
<tr>
<td>SUPER</td>
<td>IBDLD</td>
<td>S5_68137718</td>
<td>5</td>
<td>68137718</td>
<td>NA</td>
<td>2.97 x 10^{-06}</td>
</tr>
<tr>
<td></td>
<td></td>
<td>S2_193723829</td>
<td>2</td>
<td>193723829</td>
<td>NA</td>
<td>3.49 x 10^{-06}</td>
</tr>
<tr>
<td></td>
<td></td>
<td>S2_194075707</td>
<td>2</td>
<td>194075707</td>
<td>NA</td>
<td>7.72 x 10^{-06}</td>
</tr>
<tr>
<td></td>
<td></td>
<td>S2_193723781</td>
<td>2</td>
<td>193723781</td>
<td>NA</td>
<td>7.77 x 10^{-06}</td>
</tr>
<tr>
<td></td>
<td></td>
<td>S2_194923519</td>
<td>2</td>
<td>194923519</td>
<td>NA</td>
<td>8.28 x 10^{-06}</td>
</tr>
<tr>
<td></td>
<td></td>
<td>S8_3532960</td>
<td>8</td>
<td>3532960</td>
<td>NA</td>
<td>1.25 x 10^{-05}</td>
</tr>
<tr>
<td></td>
<td></td>
<td>S10_145565167</td>
<td>10</td>
<td>145565167</td>
<td>NA</td>
<td>1.71 x 10^{-05}</td>
</tr>
<tr>
<td></td>
<td></td>
<td>S2_200088348</td>
<td>2</td>
<td>200088348</td>
<td>NA</td>
<td>2.23 x 10^{-05}</td>
</tr>
<tr>
<td></td>
<td></td>
<td>S5_68026277</td>
<td>5</td>
<td>68026277</td>
<td>NA</td>
<td>2.26 x 10^{-05}</td>
</tr>
<tr>
<td></td>
<td></td>
<td>S2_195283322</td>
<td>2</td>
<td>195283322</td>
<td>NA</td>
<td>2.31 x 10^{-05}</td>
</tr>
<tr>
<td></td>
<td></td>
<td>S1_181913811</td>
<td>1</td>
<td>181913811</td>
<td>NA</td>
<td>2.36 x 10^{-05}</td>
</tr>
<tr>
<td></td>
<td></td>
<td>S5_68262516</td>
<td>5</td>
<td>68262516</td>
<td>NA</td>
<td>2.75 x 10^{-05}</td>
</tr>
<tr>
<td></td>
<td></td>
<td>S2_189709094</td>
<td>2</td>
<td>189709094</td>
<td>NA</td>
<td>4.03 x 10^{-05}</td>
</tr>
<tr>
<td></td>
<td></td>
<td>S4_192870199</td>
<td>4</td>
<td>192870199</td>
<td>NA</td>
<td>4.24 x 10^{-05}</td>
</tr>
<tr>
<td></td>
<td></td>
<td>S7_25022476</td>
<td>7</td>
<td>25022476</td>
<td>NA</td>
<td>4.86 x 10^{-05}</td>
</tr>
<tr>
<td></td>
<td></td>
<td>S1_172352901</td>
<td>1</td>
<td>172352901</td>
<td>NA</td>
<td>6.11 x 10^{-05}</td>
</tr>
<tr>
<td></td>
<td></td>
<td>S1_172724087</td>
<td>1</td>
<td>172724087</td>
<td>NA</td>
<td>6.11 x 10^{-05}</td>
</tr>
<tr>
<td></td>
<td></td>
<td>S1_184161380</td>
<td>1</td>
<td>184161380</td>
<td>NA</td>
<td>6.11 x 10^{-05}</td>
</tr>
<tr>
<td></td>
<td></td>
<td>S1_77399067</td>
<td>1</td>
<td>77399067</td>
<td>NA</td>
<td>6.11 x 10^{-05}</td>
</tr>
<tr>
<td></td>
<td></td>
<td>S4_204074328</td>
<td>4</td>
<td>204074328</td>
<td>NA</td>
<td>6.11 x 10^{-05}</td>
</tr>
<tr>
<td></td>
<td></td>
<td>S4_204250183</td>
<td>4</td>
<td>204250183</td>
<td>NA</td>
<td>6.11 x 10^{-05}</td>
</tr>
<tr>
<td></td>
<td></td>
<td>S1_87449282</td>
<td>1</td>
<td>87449282</td>
<td>NA</td>
<td>6.98 x 10^{-05}</td>
</tr>
<tr>
<td></td>
<td></td>
<td>S1_87449304</td>
<td>1</td>
<td>87449304</td>
<td>NA</td>
<td>6.98 x 10^{-05}</td>
</tr>
<tr>
<td></td>
<td></td>
<td>S7_25022476</td>
<td>7</td>
<td>25022476</td>
<td>NA</td>
<td>2.00 x 10^{-06}</td>
</tr>
<tr>
<td></td>
<td></td>
<td>S1_226374236</td>
<td>1</td>
<td>226374236</td>
<td>NA</td>
<td>3.52 x 10^{-06}</td>
</tr>
<tr>
<td></td>
<td></td>
<td>S2_193723829</td>
<td>2</td>
<td>193723829</td>
<td>NA</td>
<td>7.20 x 10^{-06}</td>
</tr>
<tr>
<td></td>
<td></td>
<td>S7_25022476</td>
<td>7</td>
<td>25022476</td>
<td>NA</td>
<td>2.00 x 10^{-06}</td>
</tr>
<tr>
<td></td>
<td></td>
<td>S1_226374236</td>
<td>1</td>
<td>226374236</td>
<td>NA</td>
<td>3.52 x 10^{-06}</td>
</tr>
<tr>
<td></td>
<td></td>
<td>S2_193723829</td>
<td>2</td>
<td>193723829</td>
<td>NA</td>
<td>7.20 x 10^{-06}</td>
</tr>
</tbody>
</table>

Both: model using the IBDLD and G matrix; bp: base pairs; NA: non-estimable in GAPIT 3 software.
Figure 4.1. Values for A) the statistic Delta K for a model-based clustering algorithm and B) the Bayesian Information Criterion (BIC) for a principal component analysis (PCA) to identify the best K within the 487 doubled haploid lines. K corresponds to the number of subpopulations. Delta K statistic detects the rate of change between successive K values, while BIC detects the fit of the model when K varies from 1 to 10. The highest delta K value corresponds to the true value for K, while the lowest BIC corresponds to the best model according to the data provided.
Figure 4.2. Quantile-quantile plots for plant height based on observed versus expected $-\log_{10}(P \text{ values})$ for the CMLM, FarmCPU, FaST-LMM, MLMM, and SUPER methods when the A) IBDLD and B) G kinship matrix were used.
Figure 4.3. Comparison of significant SNPs associated with plant height between SUPER and FarmCPU using both the VanRaden (2008) (G Matrix) and IBDLD kinship matrices.

**References**


Druet T, Gautier M (2017) A model-based approach to characterize individual inbreeding at both global and local genomic scales. Molecular ecology 26:5820-5841


Endelman JB (2011) Ridge regression and other kernels for genomic selection with R package rrBLUP. The Plant Genome 4:250-255


Patterson HD, Thompson R (1971) Recovery of Inter-Block Information when Block Sizes are Unequal. Biometrika 58:545-554


Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ, Sham PC (2007) PLINK: A tool set for whole-genome association and population-based linkage analyses. American Journal of Human Genetics 81:559-575


Appendix: Supplemental figure

Supplemental Figure 4-1. Scatter plot of the discriminant analysis of principal components (DAPC). The three populations C0 (purple), C17 (blue), C0C17 (red), are depicted inside their 95% inertia ellipses. The dots, triangles, and diamonds represent each of the DH lines within their respective population. The axes represent the first two discriminant functions, respectively. The plots of eigenvalues show the amount of genetic information retained by each of the PC axes (box on the right).
CHAPTER 5. GENERAL CONCLUSION

Lamkey and Edwards (1999) proposed a model of heterosis allowing any level of inbreeding \( (f) \) in any generation. Under their model, the mean agronomic performance of each generation was expressed as a weighted average of \( 1 - f \) times the mean of the panmictic population plus \( f \) times the mean at complete inbreeding. In doing so, five genetic parameters were described, \( \mu, \Delta^2d, \bar{p}_1\bar{p}_2d, \Delta a, \) and \( \Delta \alpha \), where \( a \) and \( d \) correspond to the additive and dominant gene effects, \( \bar{p}_1 \) and \( \bar{p}_2 \) to the average allele frequency for alleles 1 and 2, respectively, \( \alpha \) to the average effect of an allele substitution, and \( \Delta \) to the genetic divergence in allele frequency between the parents (Lamkey and Edwards 1999). The significance of each parameter to the means of the generations was evaluated by a weighted regression analysis, where the weight was the inverse of the variance of the means. Our results suggest that for grain yield (Mg ha\(^{-1}\)), plant height (cm), and ear height (cm) a single-locus model including the five genetic parameters was adequate to explain the variability observed in each generation and to predict heterosis. As expected, the effect of genetic divergence in allele frequency and inbreeding depression were significant for all traits and crosses, however, both effects are affected by the level of heterozygosity within a population. A decrease in the amount of heterozygosity compared to the common base population (i.e. an increased \( F_{ST} \)) could be caused by a decrease in the effective population size, isolation of the populations, and/or different selection pressures (Hedrick 2005), leading to an increase in the genetic divergence in allele frequency between populations because of the fixation of alternative alleles. Edwards and Lamkey (2003) concluded that by increasing the level of inbreeding within a population, a significant reduction in genetic diversity is obtained, while the variance among populations is increased. The decrease in heterozygosity within a population results in the expression of different recessive or partially...
recessive deleterious alleles that will cause a decline in the population means (Charlesworth and Charlesworth 1999). However, as the panmictic populations become inbred (increased $F_{ST}$), lower inbreeding depression will be obtained during each cycle because of a reduction in heterozygosity. Consequently, reduced heterozygosity results in low inbreeding depression and high genetic divergence between populations.

Midparent heterosis (MH) is a function of dominance and genetic divergence in allele frequencies between the parents ($4\Delta^2d$), while inbred-midparent heterosis (IMH) depends on $\Delta$, dominance, and inbreeding depression ($2\bar{p}_1\bar{p}_2d + 2\Delta^2d$). Both MH and IMH estimates were significant for all traits and crosses. For inbred line crosses (i.e. inbred-by-inbred), MH and IMH were statistically the same because of an inbred line constitutes the limit that an individual can reach when inbred to homozygosity, and therefore, the mean of the parent before and after inbreeding will be the same. During inbreeding, the panmictic populations will have a different genotype frequency, i.e. increased homozygotes at the expense of the heterozygotes, which results in changes in the population means because of reduced heterozygosity (Falconer and Mackay 1996; Lynch and Walsh 1998). However, a significant reduction in the population means will be obtained when the mean of the population before inbreeding (panmictic population) is compared to the mean at complete inbreeding. Therefore, inbred-midparent heterosis, which depends on the mean of the populations at complete inbreeding, is expected to be higher than MH in crosses involving a panmictic population because of a significant reduction in the mean of the population because of inbreeding ($2\bar{p}_1\bar{p}_2d$). In general, crosses involving inbred lines (e.g. population-by-inbred and inbred-by-inbred) showed lower midparent values and higher midparent heterosis because of a reduced level of heterozygosity in the parents. Decreasing heterozygosity (increased $F_{ST}$) results in higher heterosis because of increased
genetic divergence between parents ($\Delta$) and low inbreeding depression within parents. Hence, there is a tradeoff between inbreeding depression and heterosis, revealing that by increasing the genetic divergence in allele frequency between parents ($\Delta$) the inbreeding depression and midparent values are reduced because of the reduction in heterozygosity.

Maize breeding programs around the world have focused their efforts on the production of divergent inbred lines that when crossed have an increased agronomic performance or hybrid vigor. However, this has created a dependency of farmers to buy hybrid seed every cycle to maximize grain yield. According to Pixley and Banziger (2002), in developing countries, farmers are reluctant to use hybrid seeds because of the cost of the seed, the non-availability of hybrid seed in the region, the need to use highly technical agronomic practices, and/or the lack of adaptation of hybrid seed among others, raising the question of the advantage of hybrids versus open-pollinated or local varieties under poor farmers conditions. Several authors have reported a loss in performance caused by saving seeds from $F_1$ hybrids or planting an open-pollinated variety (Hallauer and Miranda 1988; Meghji et al. 1984; Pixley and Banziger 2002). Our results showed a loss attributed to a reduction of half the heterosis observed in the $F_1$ hybrid, with crosses involving only inbred lines (i.e. inbred by inbred) showing the higher loss in performance and crosses between panmictic populations (i.e. population-by-population) reporting the lower loss. This goes with Hallauer and Miranda (1988), who concluded that reductions in the mean performance of the $F_2$ generation are smaller when parents are open-pollinated varieties compared to cases where inbred lines are used. Equally, Lynch (1991) stated that a reduction in fitness of the $F_2$ hybrids when compared to the $F_1$ hybrids is greater when the two parents are genetically distant, which is obtained when the parents are inbred lines derived from different
populations. Finally, our results suggest that saving seed from F\textsubscript{1} hybrid results in a loss in performance attributable to half inbred-midparent heterosis or $\frac{1}{2}(2\bar{p}_1\bar{p}_2d + 2\Delta^2d)$.

Both model and empirical evidence in chapter 2 show that population divergence and inbreeding depression are the key drivers of heterosis, but that both are dependent on the level of heterozygosis within each population. Therefore, having a deeper understanding of inbreeding could lead to better predictions of heterosis. Chapter 4 focused on detecting identity-by-descent (IBD) as a measurement to account for inbreeding among inbred lines. Commonly IBD is measured by the use of pedigree information or, when marker information is available, with the use of the probability of identity-by-state (IBS) at casual loci as an approximation to the probability of IBD (Yu et al. 2006). The genetic contribution from a set of progenitors to its descendants was defined as the theoretical number of genes that trace back to those progenitors (Delannay et al. 1983; Mikel et al. 2010). Commonly, genotypic data have been used to measure the length of shared IBS segments in the genome, allowing the estimation of genetic diversity and the ancestor’s genetic contribution. However, as stated in the literature, DNA segments that are identical by descent are, by definition, identical by state (IBS), but not all IBS segments are copies coming from a common ancestor (Bernardo 2010; Falconer and Mackay 1996; Hartl et al. 1997; Walsh and Lynch 2018), and thus, an overestimated value for both genetic diversity and progenitor’s genetic contribution could be produced. As expected, our results showed a higher genetic divergence between heterotic groups than the diversity found within groups. Maize breeding programs have focused their efforts on the production of crosses between genetically divergent heterotic groups, while within-groups improvement was mainly driven by the recycling of closely related inbred lines (Mikel 2008), resulting in less genetic diversity within than between groups.
In the analysis, the inbred lines B14, B37, B73, PHG39, PHJ40, Mo17, Oh43, and PH207 were selected as key ancestors because they were known to be present in the pedigree of most of the lines used in this experiment and they have constituted some of the most recombined inbreds appearing in the derivation of other PVP lines in the U.S. (Mikel and Dudley 2006; van Heerwaarden et al. 2012). Based on a kinship coefficient computed from pedigree information (pedigree-based) or from the posterior probability of the IBD status at each locus (marker-based), nine heterotic subgroups, corresponding to the 8 key ancestors plus the Iodent Lancaster (with representative line PHZ51), were identified. The marker-based kinship has the advantage to find some hidden relationships even in the absence of pedigree information. In that regard, for the lines LH162, LH38, PHG83, PHHV4, PHJ65, PHK76, PHP60, PHW03, and Q381, from which pedigree information was obtained only for one generation, were included in one of the nine main heterotic subgroups when the marker-based distance matrix was used, while with the pedigree-based distance matrix they were included in individual groups. Our analysis using the marker-based kinship coefficient to compute a distance matrix grouped some inbred lines close to certain ancestors. For instance, the ex-PVP line PHJ65 was clustered within the PHG39 heterotic subgroup, however, based on pedigree records collected in the present study, this line was classified in the non-stiff stalk heterotic group. White et al. (2020), using a diversity estimate of 1 minus the probability of pairwise identity-by-state between 329 ex-PVP and 8 ancestors, clustered the PHJ65 inbred line within the Iodent Lancaster subgroup, while Beckett et al. (2017) using a principal component analysis (PCA) in an IBS genotypic matrix composed of 77314 SNPs, grouped the line within the non-stiff stalk, near to the line PHP60. Differences in our results with those reported by White et al. (2020) and Beckett et al. (2017) could be attributed to the lack of other public and/or proprietary lines that account for more shared segments in the
genome with the PHJ65 line and/or to the use of a genome-wide estimate of IBD instead of a genome-wide estimate of identity by state.

A high correlation ($r = 0.86$) between pedigree-based and marker-based kinship estimators was observed, suggesting that recent inbreeding has the biggest contribution to the coancestry between individuals (Toro et al. 2002). Differences in the estimators of coancestry between the pedigree-based and marker-based kinships could be explained as a consequence of some population-specific parameters (e.g. effective population size ($N_e$), selection, or random Mendelian sampling) and to some variations in the estimation of the relationship, with pedigree depth and the number of markers affecting directly how close the estimator will be from the true value. With small effective population sizes ($N_e$) a higher influence of the level of linkage disequilibrium (LD) in a population will be produced because of the increase in the probability of alleles identical by descent (IBD) between individuals (Crow and Kimura 1970; Hill and Robertson 1968; Hill 1981; Laurie-Ahlberg and Weir 1979; Sved 1968; Wang 2016). Consequently, with high LD the probability of finding blocks of loci where all alleles are either identical or non-identical by descent is increased, affecting the kinship when computed with pedigree or marker data (Sved 1968; Wang 2016; Weir et al. 1980). Marker-based estimates will be affected by the number of available SNPs and the level of LD, suggesting that with larger $N_e$ more markers are required in order to capture the relationship between individuals (Wang 2016). Similarly, under selection, changes in allele frequencies are produced because only individuals with superior performance for the trait will contribute with alleles to the next generation (Albrechtsen et al. 2010), increasing the probability that multiple individuals inherit identical alleles. Finally, because of random Mendelian sampling, the kinship coefficient of two individuals with the same pedigree will be different when computed with markers, while when
using pedigree information both will have the same kinship because of common records. Therefore, marker-based kinship can account for changes in allele frequencies caused by Mendelian sampling, while measuring the effect of selection on mendelian sampling at the same time.

Both marker-based and pedigree-based kinship coefficients provides the probability IBD relative to a declared reference population. Marker-based kinship coefficient depends on the allele frequency of the declared reference population and for that, the current sample of individuals is used as the reference population from where allele frequencies are estimated (Wang 2014). Therefore, if the genotyped individuals are a representative sample of the reference population, unbiased estimates for kinship will be produced, however, when the number of markers is low, marker-based kinship coefficients are expected to be biased (Wang 2016). Both marker-based and pedigree-based kinship coefficients are expected to be close to each other when accurate and ample pedigree information is available, with the former being able to capture the mendelian sampling. In general, both pedigree- and marker-based kinship coefficients rely on the definition of a reference population, however, the latter can shift the reference to be the population of sampled individuals.

Our results provide a high-resolution data analysis in the identification of regions in the genome inherited from key ancestors. As expected, there was a prevalence of large IBD segments inherited from the key ancestor lines. Because of selection, large preserved regions in the genome could be inherited as a block from the parents (Hill and Robertson 1968; Jordan et al. 2004; Robertson 1961; Thompson 2013), increasing the inheritance of alleles that are in LD (Albrechtsen et al. 2010; Hospital and Chevalet 1993). For long IBD segments, an increased probability of alleles linked in a repulsion phase, with unfavorable and favorable alleles inherited
as a block (i.e. hitchhiking effect), is produced, decreasing the total genetic diversity within the population (Hospital and Chevalet 1993). This hitchhiking will create an increase in genetic drift and a significant decrease in the effective population size (Barton 2000; Smith and Haigh 1974). Therefore, the higher lengths of IBD segments found in the SSS background are evidence of small \( N_e \) and the strong or recent selection for this group, which has relied mainly on Iowa Stiff Stalk Synthetic derived lines for the accumulation of favorable alleles.

The testing of associations between genetic variants and the phenotype of a trait across the whole genome has been referred to as “Genome-Wide Association Studies” or GWAS (Hirschhorn et al. 2002; Liu 2014). This type of analysis has led to the identification of associations for several traits and crops. GWAS analysis is dependent on the ability to screen hundreds or thousands of individuals for a large number of polymorphic markers, such that the probability of finding markers associated or correlated with the causal allele increases (Hirschhorn and Daly 2005). Several methods have been implemented during the time for GWAS, with the mixed linear models (MLM) constituting the most used and accepted method because of its ability to incorporate cryptic relatedness and population structure (Aulchenko et al. 2007; Li and Zhu 2013; Mei and Wang 2016; Yu et al. 2006; Zhang et al. 2010). However, direct implementation of a mixed linear model to account for associations becomes impossible for large number of individuals (Zhang et al. 2010). Therefore, several approaches relying on grouping the individuals based on identity-by-descent, fitting a two-step approach to solving the normal equations, and/or using spectral decomposition to reduce dimensionality have been proposed (Aulchenko et al. 2007; Kang et al. 2008; Lippert et al. 2011; Liu et al. 2016; Segura et al. 2012; Zhang et al. 2010). We test several methodologies in a real data example using 487 double haploid lines for plant height. Our results showed that SUPER identify several moderately
significant SNPs. The other methods, CMLM, FaST-LMM, and MLMM were not able to detect any causative allele associated with plant height. Using FarmCPU five SNPs were found to be associated with plant height. The strongest association was found for SNP S7_25022476, which has been reported to be associated with the Lazarus 1 (LAZ1) protein (gene Zm00001d008266) involved in vacuolar transport, brassinosteroids signaling, and programmed cell death signaling pathways (Liu et al. 2020; Liu et al. 2018; Malinovsky et al. 2010). Similarly, SNP S8_3532960 was near the gene Zm00001d008266, which is involved in auxin catabolism, one of the major plant regulators involved in plant growth and plant architecture (Gallavotti 2013; Tognolli et al. 2002). SNP marker S2_193723829 was near the gene Zm00001d005951, which is associated with growth regulators in the response to water deficit (Himmelbach et al. 2002; Söderman et al. 1999). Finally, the SNPs S1_226374236 and S9_133204506 were near genes involved in membrane solute transport in the cell and chloroplast accumulation or movement in response to environmental light conditions, respectively, and for that were not directly associated with plant height (Himmelbach et al. 2002; Kodama et al. 2010; Luesse et al. 2006; Söderman et al. 1999). Therefore, three candidate genes Zm00001d008266, Zm00001d008266, and Zm00001d005951 located on chromosome 7, 8, and 2, respectively, were selected in the experiment for being associated with plant height.

Finally, the identification of IBD segments through a high-resolution analysis using a Hidden Markov model framework, allows the estimation of the identity-by-state probability between or within individuals, instead of the approximation to IBD from the detection of identity-by-state (IBS) regions (Yu et al. 2006). In the real data example the use of a relationship matrix computed from IBD probabilities does not improve the estimation of associations for plant height, however, we hypothesize that for crops where not sufficient pedigree information is
available or when the pedigrees correspond to unrelated individuals, an HMM could help in the identification of hidden relatedness resulting in an improvement in the identification of associations between the phenotype and causative locus. Similarly, the use of a HMM approach to identify IBD segments, could result in a better classification of individuals within heterotic groups, helping breeders to identify and exploit different heterotic patterns.
REFERENCES


Baker R (1984) Some of the open pollinated varieties that contributed the most to modern hybrid corn. Illinois Corn Breeding School, pp 1-19


Cockerham CC (1954) An extension of the concept of partitioning hereditary variance for analysis of covariances among relatives when epistasis is present. Genetics 859


East EM (1908) Inbreeding in corn. Rep Conn Agric Exp Sta, pp 419-428


Hill W, Robertson A (1968) Linkage disequilibrium in finite populations. Theoretical and applied genetics 38:226-231


Jacquard A (1975) Inbreeding: one word, several meanings. Theoretical Population Biology 7:338-363


Jenkins MT (1936) Corn Improvement. Yearbook of Agriculture:455-522

Jones DF (1917) Dominance of linked factors as a means of accounting for heterosis. Genetics 2:466-479


Laurie-Ahlberg CC, Weir BS (1979) Allozymic Variation and Linkage Disequilibrium in Some Laboratory Populations of DROSOPHILA MELANOGASTER. Genetics 92:1295


by testcross progenies of near-isogenic lines reveals a significant role of epistasis. Genetics 177:1827-1837


Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ, Sham PC (2007) PLINK: A tool set for whole-genome association and population-based linkage analyses. American Journal of Human Genetics 81:559-575


Rudan I, Campbell H (2004) Five reasons why inbreeding may have considerable effect on post-reproductive human health. Collegium antropologicum 28:943-950


Shull GH (1908) The composition of a field of maize. Journal of Heredity:296-301


Sprague GF (1955) Corn and corn improvement. LWW


Sved JA (1971) Linkage disequilibrium and homozygosity of chromosome segments in finite populations. Theoretical Population Biology 2:125-141


Wang J (2016) Pedigrees or markers: Which are better in estimating relatedness and inbreeding coefficient? Theoretical population biology 107:4-13


Wright S (1921) Systems of mating. II. The effects of inbreeding on the genetic composition of a population. Genetics 6:124-143

