CHAPTER 2

LITERATURE REVIEW

2.1 Maize is an important crop for genetic analysis

Maize is a member of the grass family, Poaceae. This family represents a range of genome and structural complexity ranging from diploid species with a genome size of 415 Mb in rice to 16,000 Mb in hexaploid wheat. Maize is diploid (2n = 20). Its chromosomes contain 2.5 billion base pairs and it lies somewhere in the middle of grass genome size and complexity (Gaut *et al*., 2000). Maize has been a major focus of biotechnology research because of its economic importance and naturally occurring high polymorphism.

Maize is cultivated worldwide, at latitudes varying from the equator to slightly above 50 degrees north and south, from sea level to over 3000 meters elevation, in cool and hot climates, and with growing cycles ranging from 3 to 13 months (CIMMYT, 2000). Of the major grain crops, maize has the largest total annual grain production in the world (590.5 million metric tons, mmt) followed by wheat (567.7 mmt) and rice (380.3 mmt) and its average yield per hectare (4.3 t) is more than 60% higher than that of either wheat or rice (FAPRI, 2003). In addition to its direct use as food and feed, maize also has various industrial processing products, such as wet milling (e.g. starch, and oil), dry milling (e.g. meal and flour), and fermentation and distillation (e.g. alcohol and whisky).

Maize is one of the crop species with the highest level of molecular polymorphism. For certain loci, over 5% nucleotide diversity has been reported (Henry and Damerval, 1997). Nucleotide diversity is measured as the average sequence divergence between any two individuals for a given locus (Buckler and Thornsberry,

2002). Maize is also used as model to study genome size evolution because of its polyploidy origin and the abundance of transposons (Bennetzen and Devos, 2002). Transposons are segments of DNA that can move around to different positions in the genome. In the process, they may cause rearrangements or insertion and deletions and thereby change the amount of DNA in the genome. The use of transposons to isolate genes has made it possible to overcome the difficulty of working with a large and mostly repetitive genome. Because of this, maize is used as one of the model crops for studying the biology of cereals, which provide over 70% of the total human caloric intake worldwide (FAPRI, 2003).

2.1.1 Diversity in maize

Zea is the genus for maize (*Zea mays ssp. mays*) and its wild relatives, teosinte. There are four species (*Zea diploperennis*, *Zea perennis*, *Zea luxurians*, and *Zea mays*) in this genus and they are all native to Mexico and Central America (Doebley and Iltis, 1980). Four sub-specific taxa have been recognized in *Z. mays*, namely, *Z. mays ssp. huehuetenangensis*, *Z. mays ssp. mays*, *Z. mays ssp. mexicana*, and *Z. mays ssp. parviglumis*. Among the three wild subspecies of *Zea mays*, *Z mays ssp. parviglumis* is thought to be the progenitor of cultivated maize (Doebley *et al.,* 1984; Wang, *et al*., 1999). Similarly, using microsatellite-based phylogenetic analysis of 264 of maize and its progenitor, (Matsuoka *et al*., 2001) showed that maize was domesticated from *spp. parviglumis* in southern Mexico about 9,000 years ago.

Genetic diversity studies in maize have shown that maize is highly variable both

within and across populations. Sequencing of the *adh1* locus in maize, in *Z. mays* ssp. *parviglumis* (the maize progenitor), and in *Zea luxurians* (a distant maize relative), showed that maize retained 77% of the diversity of *parviglumis* and has more diversity than *Z. luxurians* (Eyre-Walker *et al.*, 1998). Maize molecular diversity is roughly 3 to 10 fold higher than that of other domesticated grass crops (Buckler *et al.*, 2001). This is probably the result of several factors: (1) the diversity of environments, culture, production system and the type of consumption of maize (Aguirre *et al*., 1998), which facilitated the development of different maize types throughout the world; (2) the high level of out-crossing. Maize is a monoecious crop with male and female reproductive parts that are physically separated, which facilitates out-crossing. This favors continuous gene exchange between neighboring plants and in some cases, with their wild relatives; (3) the existence of chromosomal duplications. In maize chromosomal duplications are extensive, which provides new mutational opportunities for creating greater phenotypic variability (Helentjaris *et al*., 1988); and (4) the presence of transposons and retrotransposon elements (Bennetzen and Devos, 2002). Breeders rely on the natural diversity found within crop species for selection and improvement of qualitative and quantitative traits. As a result, maize yields have increased to 55 fold higher than its progenitor (Buckler *et al*., 2001).

2.1.2 Maize genome evolution

Gene or whole genome duplications (polyplodization), insertions of viral DNA, microsatellite and heterochromatin expansions, and transposon insertions all add to nuclear genomes (Bennetzen and Devos, 2002). The maize genome contains extensive

chromosomal duplication and repetitive DNA. Most repetitive DNA in the maize genome comprises retrotransposon elements (a mobile segment of DNA, which uses RNA as a template for replication), and these repetitive DNA comprise 50% of the genome (Gaut *et al*., 2000). Repetitive DNA is defined as DNA with more than 100 copies per genome. The repetitive DNA of maize can further be categorized as 20% highly repetitive (over 800,000 copies per genome) and 40% middle repetitive (over 1000 copies per genome; Hake and Walbot, 1980).

Polyploidization is a major force in plant genome evolution. It has been estimated that 50-70% of flowering plants have experienced chromosome doubling at least once in their evolutionary history (Wendel, 2000). Restriction fragment length polymorphism mapping studies have shown that many markers map to two or more chromosomal locations (Helentjaris *et al*., 1988). Ahn and Tanksley (1993) reported that 72% of the single-copy rice genes are duplicated in the maize genome. Gaut (2001) found that 60-82% of the maize genome is statistically significant for colinearity (shared markers in shared order) and nearly a third of the genome may be even in multiple copies. These findings of extensive chromosomal duplication in maize have been interpreted as evidence for a polypolid origin of the genome (Helentjaris *et al*., 1988).

2.1.3 Wild relatives of maize

The genus *Zea* consists of four species of which only *Zea mays* ssp. *mays* L. is economically important. The other *Zea* species, referred to as teosintes, are wild grasses (Doebley, 1990). These species, mostly perennials, contain a number of useful

genes. Attempts to transfer apomixis (asexual reproduction of a plant through seed) genes from *Tripsacum* to maize have been pursued for a number of years (Leblanc *et al*., 1995; Grossniklaus *et al*., 1998), and consequently patents on apomictic maize have been published (Savidan *et al*., 1998; Eubanks, 2000). Apomixis may be of great significance to the maize growing world. In developing countries, many farmers cannot take advantage of hybrid technology because hybrid seed is either unavailable or unaffordable. If apomixis could be applied in maize, farmers would have the opportunity to recycle hybrid seed from generation to generation, thereby avoiding the cost of buying new seed each season.

In Africa, the parasitic weed *Striga spp*. is a significant pest of maize and sorghum. Little resistance has been found within cultivated maize (Hoisington *et al*., 1999). A potential valuable source of resistance to *Striga hermonthica* may lie in the genetic potential of a wild relative of maize (Tanksley and McCouch, 1997). In conclusion, wild relatives of maize represent significant untapped genetic resources for the improvement of maize.

2.2 Methods for assessing genetic variation

Knowledge of the genetic variation in crop collections is essential for their efficient use in breeding programs, as well as to establish new collections and conservation strategies. Exploiting natural variation is very important for several reasons: (1) genetic variability in crops is advantageous, because it allows the crop to adapt to new biotic and abiotic stresses, (2) many landraces and wild relatives of crop plants contain desirable genes that

confer resistance to pests and diseases, and control quality traits. For example, approximately, 40-80% of the yield gain in maize, wheat and barely has been obtained from genetic improvements of these crops (Evans and Evans, 1993; Hallauer and Miranda, 1988). Consequently, large numbers of varieties are being collected around the world in an effort to conserve the genetic variation and provide access to valuable material for plant breeders. The international center for maize and wheat improvement maintains 17,000 maize accessions (CIMMYT, 2000). As the number of accessions increases, it becomes more difficult to avoid the inclusion of duplicate or at least very similar accessions. Evaluation of numerous, highly similar accessions not only wastes plant breeding resources but likely reduced the chance of identifying the truly unique and valuable accessions. In addition, field evaluation of the whole collection for a variety of traits is difficult because it is laborious and time consuming.

To evaluate and utilize these collections, it is necessary to identify a smaller subset or core collection that likely represents most of the genetic variation in the entire collection. Brown (1989a & b) suggested that at least 70% of the alleles present in the entire collection would be represented in a core collection comprised of at least 10% of the accessions, provided that the selection of the core collection is carried out systematically to capture most of the diversity. To assemble a core collection, numerical methods may be useful for directing the selection of accessions. The data could be agro-morphological performance, pedigree relationships or molecular marker information (genetic diversity).

2.2.1 Morphology and pedigree data

Morphological traits were among the earliest genetic markers used in germplasm management (Stanton *et al*., 1994) but they have a number of limitations, including low polymorphism, low heritability, late expression, and vulnerability to environmental influences (Smith and Smith, 1992), which in turn may affect the estimation of genetic relationships. Therefore, to be useful, morphological measurements should be accomplished in replicated trials. This may be expensive and time consuming. However, if the traits are highly heritable, morphological markers are one of the choices for diversity studies because the inheritance of the marker can be monitored visually without specialized biochemical or molecular techniques. In maize, qualitative and quantitative traits have been used to establish core collection (Taba *et al*., 1998) and to study phenotypic diversity (Alika *et al*., 1993; Lucchin *et al*. 2003).

To quantify the relationship based on pedigree information, Malecot (1948) presented the coefficient of co-ancestry (*f*), also known as the kinship coefficient or the coefficient of parentage. Pedigrees of varieties are defined as a complete record of relationships traced back to landraces and wild relatives. This measure estimates the probability that two randomly drawn, homologous genes (alleles) from each of two individuals are identical by descent. The measure based on Mendelian inheritance and probability is calculated under several assumptions: (1) the absence of selection, mutation, migration and drift, (2) regular diploid meiosis, and (3) no relationship for individuals without a verified common ancestor (Melchinger, 1993).

Several common features of plant breeding programs cause departures from these assumptions because of (1) intense selection, (2) drift due to small sample sizes, and (3) unknown or incorrect pedigree records (Messmer *et al*., 1993). Despite this, it has been widely used in self-pollinated crop species such as barley, wheat, soybean and peanut to examine the level of genetic diversity and identify major groupings of related cultivars (Martin *et al.,* 1991). Accurate estimation of genetic similarity by co-ancestry requires reliable and detailed pedigree records. However, for many maize inbreds and their progenitors, pedigree records tracing back more than two generations are rare or incomplete and calculation of co-ancestry for maize is not feasible (Messmer *et al*., 1993).

2.2.2 Molecular markers

Molecular markers are useful tools for assessing genetic diversity among germplasm compared with morphology and pedigree information because they are not affected by environmental factors. A molecular marker is a variant of DNA or a protein which can be detected and whose inheritance can be monitored reliably (Jones *et al*., 1997). Compared with morphological and pedigree information, molecular markers reveal differences among genotypes at the DNA level and thus provide a more direct, reliable and efficient tool for germplasm conservation and management. As a result, researchers are adopting molecular markers as valuable tools for genetic diversity studies in many crops.

In past decades, marker systems such as Restriction Fragment Length Polymorphisms (RFLPs), Randomly Amplified Polymorphic DNA (RAPDs, Welsh and McClell, 1990; Williams *et al*., 1990), Amplified Fragment Length Polymorphisms (AFLPs, Vos *et al., 1995*), microsatellites or Simple Sequence Repeats (SSRs, Tautz, 1989), single nucleotide polymorphisms (SNPs) and others have been developed and applied to a range of crop species. In general two basic types of marker systems are available (1) those that rely on hybridization between a probe and homologous DNA segment within the genome, and (2) those that use polymerase chain reaction (PCR) to exponentially amplify genome segments between arbitrary or specific oligonucleotide premier sites (reviewed by Karp *et al*., 1996; Jones *et al.,* 1997; Kumar, 1999).

RFLP analysis was one of the first techniques to be used widely to detect variation at sequence level. It examines the variation in size of specific DNA fragments following digestion with a restriction enzyme. RFLP is co-dominant and hence we can distinguish the heterozygote from homozygote individual (Helentjaris *et al*., 1985). RFLPs have been used in maize to investigate pedigree relationships among inbreds and to assign them to heterotic groups (Melchinger, 1993; Dubreuil *et al*., 1996), investigating genetic diversity and relationships (Pejic *et al.,* 1998; Rebourg *et al*., 2001; Gauthier *et al*., 2002) and for the development of genetic maps (Helentjaris *et al*., 1986; Gardiner *et al.,* 1993; Coe *et al*., 1995). However, a disadvantage of RFLPs is that large quantities of DNA are required, which limits the number of marker assays that can be performed on an individual plant and the technique is difficult to automate. As a result, it is increasingly substituted by other marker techniques based on the polymerase chain reaction (PCR) such as RAPDs, AFLPs and SSRs (Jones *et al*.,

1997), because these systems allow essentially unlimited marker assays per individual.

RAPD technology is another procedure used to detect nucleotide sequence variation This PCR-based technique requires neither cloning nor sequencing of DNA. It can detect several loci simultaneously. Short (8-12 bp) arbitrary primer sequences are used to amplify DNA, usually resulting in presence/absence of polymorphisms. Moeller and Schaal (1999) studied genetic variation among 15 Native American maize accessions and found an average polymorphism of 70.7% for the 11 primers analyzed. Although, RAPD analysis is easy, inexpensive and fast, its reproducibility is problematic due to the short primers being easily affected by low annealing temperatures (Demeke *et al.,* 1997; Karp *et al*., 1997).

The AFLP technique combines the restriction site recognition element of RFLP analysis with the exponential amplification aspects of PCR-based markers. It is similar to RAPD analysis, but the primer consists of a longer fixed portion (about 15 bp) and a short (2-4 bp) random portion. The fixed portion gives the primer stability (and hence enhances repeatability) and the random portion allows it to detect a specific subset of loci. Other advantages of the AFLP technique include: (i) no sequence information is required, (ii) the PCR technique is fast, and (iii) it has a very high multiplex ratio (up to 100 genetic loci may be simultaneously analyzed per experiment). This makes it suitable for large-scale genetic diversity studies in crop species. However, AFLP and RAPD are dominant markers, which prohibits the identification of heterozygote from homozygote. This makes AFLP and RAPD less

informative than other co-dominant markers (e.g. RFLP and SSR markers).

In maize, AFLP markers have been employed (i) to investigate the relationship between genetic distances and hybrid performance for yield (Ajmone-Marsan *et al.,* 1998; Melchinger *et al.,* 1998), (ii) to study the genetic similarity of inbreds (Pejic *et al*., 1998; Lubberstedt *et al.,* 2000; Vuylsteke *et al.,* 2000b), (iii) to identify chromosomal regions involved in hybrid performance and heterosis (Vuylsteke *et al*., 2000a) and for construction of genetic linkage map (Vuylsteke *et al.,* 1999).

SSR markers has been a marker system of choice for population genetic studies, because it combines many desirable properties including co-dominance, high variability, rapid and simple assays, and uniform genome coverage (Powell *et al*., 1996). In addition, automated PCR-based technique, which enables high-throughput data collection and good analytical resolution at a low cost, has been developed for microsatellites (Mitchell *et al*., 1997). Because of these qualities, it is frequently applied in genetic diversity studies in maize inbred lines and out-bred populations (Senior *et al*., 1998; Matsuoka *et al*., 2002; Warburton *et al*., 2002; Pinto *et al*., 2003) and to identify and map quantitative trait loci (QTLs) for grain yield and yield components in maize (Thornsberry *et al*., 2001; Mohammadi *et al*., 2002).

One advantage of microsatellite analysis is the large number of polymorphisms that the method reveals per locus, increasing the informativeness of SSR markers. A locus in maize can have up to 16 alleles (Warburton *et al*., 2002). The high allelic diversity is a product of their high rate of stepwise mutation due to replication slippage

(Levinson and Gutman, 1987). The stepwise mutation model assumes that alleles mutate back and forth by small number of repeats, and thus the same allelic state are created repeatedly over time. An alternative model is the infinite alleles model (Ohta and Kimura, 1973), which assumes that each mutation creates a new allele in the populations. Matsuoka *et al*. (2002) reported that out of 46 maize microsatellite loci analyzed on all the diploids of *Zea* and 101 maize inbreds, only two followed stepwise allelic distribution, while four were nearly stepwise, 13 mixed (stepwise and continuous), eight nearly continuous and 19 continuous.

In recent years, SNPs (single base pair positions at which different sequence alternatives exist between two individuals) have become an increasingly important class of molecular marker due to its abundance (present in all parts of the genome) and amenability to fully automated genotyping (micro-array procedures have been developed for automatically scoring hundreds of SNP loci simultaneously at a low cost per sample). A high throughput assay for the detection and validation of SNPs were developed in maize. These techniques allow the rapid production of valuable information on the genetic relationships among maize varieties.

The marker system of choice depends on the objective of the study, skills and facilities available in the laboratory. The relative advantages and disadvantages of these techniques are summarized in Table 2.1.

Table 2.1 Comparison of the most common used marker systems in plant breeding

Characteristics	RFLPs	RAPDs	AFLPs	SSRs	SNPs
DNA required (µg)	10	0.02	0.5-1.0	0.05	0.05
DNA quality	High	High	Moderate	Moderate	High
PCR-based	No	Yes	Yes	Yes	Yes
Level of polymorphism	High	Moderate	High	Very high	High
Ease of use	Not easy	Easy	Easy	Easy	Easy
Amenable to automation	Low	Moderate	High	High	High
Reproducibility	High	Unreliable	High	High	High
Development cost	Low	Low	Moderate	High	High
Cost per analysis	High	Low	Moderate	Low	Low

2.3 Use of pooled DNA samples in the study of genetic variation

Genetic variation is important in the process of crop improvement and is also the basis of genetic fingerprinting. Accordingly, there has been an interest in studying genetic variation through the introduction of different DNA-based marker techniques. Although most marker techniques are relatively simple and rapid, the large number of individual plants that need to be processed may limit the application of DNA-based marker analysis of entire germplasm collections. Because DNA-based marker analysis is quite expensive, the total cost of any such project usually limits the number of genotypes that can be analyzed. The problem is most acute for out-crossing species. Crossa *et al*. (1993) showed that for out-bred maize varieties, with 48 individuals per population, with 5 loci and 5 alleles per locus, there is a 95% probability of detecting all alleles with a frequency of 0.05 or greater. If only 24 individuals are analyzed, only alleles with frequency of 0.12 or greater can be detected at this level of

probability. Hence, genotyping of open-pollinated species using DNA-based markers is 24 to 48 times more expensive than that of self-pollinated species.

One approach to overcome this limitation is to analyze one, or several, bulked samples per population, rather than individual plants. Bulking of DNA samples not only drastically reduce the number of samples that need to be processed, but also results in dilution of rare alleles (Michelmore *et al*., 1991), and therefore simplifies the marker profile of an individual population. Bulking strategies provide a means for large-scale diversity analysis in out crossing plant species. Guthridge *et al*. (2001) in their genetic diversity analysis of perennial ryegrass using AFLP recommended pooling of DNA samples in order to ensure equivalent representation in the AFLP template. Furthermore, parallel studies in white clover (Kolliker *et al*., 2001) demonstrated that bulking at the leaf stage is effective in producing representative profiles of varieties. Both studies have found that bulks of 20 individuals for perennial grass were adequate to study within and between population variations. Similarly, using bulk RFLP methods (two 15-plant bulks per population), Rebourg *et al*. (2001) described the genetic relationships among the 131 European maize populations. All of these results indicate the feasibility of bulking DNA or leaf samples from 15-30 individuals per accession/population as a cost efficient and effective means of characterizing open-pollinated crops.

2.4 Correlation between phenotypic and molecular markers distance

The use of different molecular markers to evaluate genetic diversity may reveal

different patterns of variation due to inherent differences among marker systems. Differences detected by molecular markers are not necessarily correlated with phenotypic variation, because molecular markers can potentially cover the entire genome (coding as well as non-coding regions), and most of the genome is composed of non-coding DNA, it is reasonable that the majority of differences detected by molecular markers are from non-coding regions, while phenotypic differences are brought to specific genes or coding regions. Therefore, the combination of morphological and molecular information is required to describe correctly the relationships among genotypes.

Different researchers have studied the relationship between marker and morphological information. Theoretical results of Burstin and Charcosset (1997) suggested that the relationship between morphological and marker distance is most likely triangular. This means close genetic relationships correspond with close morphological relationships, whereas distant genetic relationships can correspond with both close and distant morphological relationships. In many cases, the correlation between distances based on morphology and molecular markers are not straightforward to interpret. Consequently, a combination of morphological and molecular analyses may be the most useful to understand all aspects of genetic variation within a species or populations. Based on this approach, Rebourg *et al*. (2001) classified European maize populations into six major groups that were consistent with the origin of the populations.

2.5 Statistical measures for assessing genetic diversity

Classifying genotypes into clusters based on molecular markers and agro-morphological traits for studying genetic and phenotype diversity is a common practice. Once the morphological traits or molecular profiles have been generated, various genetic distance measures have been proposed. Genetic distance is defined as any quantitative measure of genetic difference calculated between individuals, populations or species (Beaumont *et al*., 1998).

2.5.1 Types of distance measures

Genetic distance between individuals can be calculated by various statistical measures depending on the type of data. The Euclidean distance (straight-line) and squared Euclidean distance are two commonly used measures of dissimilarity between individuals based on morphological data. Dissimilarity coefficients estimate the distance or unlikeness of two individuals, the larger the values the more different the two individuals. While similarity indices measures the amount of closeness between two individuals, the larger the value the more similar the two individuals. For molecular data, the commonly used measures of genetic similarity (GS) are (i) Nei and Li's (1979) coefficient (GS_{NL}), (ii) Jaccard's (1908) coefficient (GS_J), (iii) Simple matching coefficient (GS_{SM}) (Sokal and Michener, 1958), and (iv) Modified Roger's distance (GS_{MR}). Genetic distances between two individuals *i* and *j* determined by these measures can be obtained as follows:

$$GD_{NL} = 1-[2N_{11}/(2N_{11}+N_{10}+N_{01})]$$

$GD_J = 1-[N_{11}/(N_{11}+N_{10}+N_{01})]$

$GD_{SM} = 1-[(N_{11}+N_{00})/(N_{11}+N_{10}+N_{01}+N_{00})$

$GD_{MR} = 1-[(N_{11}+N_{10})/2N]^{0.5}$

Where N_{11} is the number of bands/alleles present in both individuals; N_{00} is the number of bands absent in both individuals; N_{10} is the number of bands present only in individual *i*; N_{01} is the number of bands present only in individual *j*; and N represents the total number of bands. The GS_{NL} formula excludes bands absent in both individuals, which cannot be necessarily attributed to a common cause. In contrast, GD_{SM} gives equal weight to mismatches and matches of bands in both individuals (Link *et al*., 1995).

Most researchers use more than one measure of genetic distance to analyze a given data set. In such case, it is important to test the correlation between matrices derived from different distance measures. One such test is the Mantel test (Mantel, 1967). The Mantel test can be performed on dissimilarity or similarity matrixes and can be applied to different types of variables. This is especially important for the analysis of genetic diversity, where various types of data sets (e.g. morphological, biochemical or molecular markers) may be used to assess the relationships among individuals. The significance of correlation can be tested via permutation procedure (Manly, 1991).

2.5.2 Multivariate methods

Multivariate techniques, which simultaneously analyze multiple measurements on

each individual under study, are widely used in analysis of genetic diversity in morphological and molecular marker data. Among these methods, cluster and principal components analyses are most commonly used. Cluster analysis refers to a group of multivariate techniques, whose primary purpose is to group individuals based on the characteristics they possess so that individuals with similar descriptions are mathematically gathered into the same cluster (Hair *et al*., 1995). The resulting cluster of individuals should then exhibit high within cluster homogeneity and high between cluster heterogeneity. Principal component analysis (PCA) is defined as a method of data reduction to clarify the relationships between two or more characters and to divide the total variance of the original characters into a limited number of uncorrelated new variables (Wiley, 1981). PCA can be used to drive a two-dimensional scatter plot of individuals, such that the geometrical distance among individuals in the plot reflect the genetic distances among them with minimal distortion. Aggregations of individuals in such a plot will reveal sets of genetically similar individuals (Warburton and Crossa, 2000).

Principal Coordinate Analysis (PCO) is another data reduction method commonly used by breeders and geneticists. The goal of PCO is to permit the positioning of objects in a space of reduced dimensionality while preserving their distance relationships as much as possible. The value of PCO is that it permits the use of all types of variables, provided that a coefficient of appropriate type has been used to compute the resemblance half-matrix. PCO differs from PCA in the way in which the data swarm is constructed to begin with. In PCO, the points are not plotted in an *s*-dimensional coordinate frame. Instead, dissimilarities are calculated between every

possible pair of objects, and the points plotted in such a way as to make the distance between every pair of points as nearly as possible equal to their dissimilarity. One could argue that PCO is necessarily inferior to PCA because in PCA each point is placed exactly where it ought to be, whereas in PCO each point is only approximated based on a best-fit model of the dissimilarities.

2.5.3 Clustering methods

There are broadly two types of clustering methods: (1) distance-based methods, in which a pair-wise distance matrix is used as input for clustering analysis (Johnson and Wichern, 1992). The result can be visualized as a tree or dendrogram in which clusters may be identified, and (2) model-based methods in which observations from each cluster are assumed to be random draws from some parametric model, and inference about parameters corresponding to each cluster and cluster memberships of each individual are performed jointly using maximum-likelihood or Bayesian methods.

Distance-based methods can be further categorized into hierarchical and non-hierarchical. Hierarchical clustering is performed by a serious of successive mergers (agglomerative) of groups of individuals. The most similar individuals are first grouped and these initial groups are merged according to their similarities. The Unweighted Paired Group Method using Arithmetic averages (UPGMA, Sneath and Sokal, 1973) and Ward minimum variance methods (Ward, 1963) are the most commonly used agglomerative hierarchical clustering methods. The non-hierarchical

clustering procedures, also known as K-means clustering methods, are based on sequential threshold approaches for assigning individuals to specific clusters after the number of clusters to be formed is specified (Everitt, 1980). This method is rarely used for genetic diversity study because the lack of prior information about the optimal number of clusters that is required for accurate assignment of individuals.

2.5.4 Partitioning of variation

When a set of populations is investigated, the amount of genetic variability can be expressed at different hierarchical levels, e.g., between agroecologies, between populations within agroecologies and within populations. For molecular data, the analysis of molecular variance (AMOVA, Excoffier *et al*., 1992) has been widely used (Warburton *et al*., 2002; Reif *et al*., 2003) for estimation of the variance components among and within the group. AMOVA is based on squared Euclidean distances among individuals, and assumes that the studied populations are in Hardy-Weinberg equilibrium. A similar method is known as analysis of distance (AOD, Van-Eeuwijk and Baril, 2001), which can be applied for any distance matrix be it Euclidean or not, and for any type of marker (morphological, molecular or a mixture) using the following formula:

$$d^2_{gi;g'i'} = \sum_{m=1}^{M} \left(x_{mgi} - x_{mg'i'}\right)^2,$$

Where d^2 is the distance between an individual *i* in group *g* and an individual *i'* in group *g'* for marker m (from 1 to *M).* Accordingly, the squared distance is the sum of the squared differences between individual accessions over all variables. The total variation (V_T) can be shown to be equal to the sum of all squared pair-wise distances between individual accessions (over all groups), divided by the total number of accessions

(overall groups). The within group variation (V_W) is the sum over groups of the sum of squared pair-wise distances within a group divided by the group size. The between group variation (V_B) can be obtained by subtraction, $V_B = V_T - V_W$.

2.6 Gene mapping/tagging

The identification and use of major genes controlling quantitative traits (QTLs) have been a major focus in maize breeding. For simply inherited traits, a difference between parents in one or two genes may explain nearly 100% of the differences among the progeny. However, many agriculturally important traits such as yield, quality and plant height show continuous variation among individuals and such traits are termed quantitative traits. In maize, many quantitative traits have been extensively investigated using conventional biometric approaches (Hallauer and Miranda, 1988). The concept of detecting QTLs was developed more than 80 years ago (Sax, 1923). However, the development of genetic markers has had great impact in the field of quantitative genetics, mainly for identifying the chromosomal segments or individual genes underlying a quantitative trait (Kumar, 1999; Stuber *et al*., 1999, Bouchez *et al*., 2002).

Molecular markers also permit plant breeders to correctly map or place the various genes that condition complex agronomic traits. Mapping is putting markers in order, indicating the relative genetic distances between them, and assigning them to their linkage groups on the basis of the recombination values from all their pair-wise combinations (Jones *et al*., 1997). Genetic mapping is essential for effective

manipulation of important genes, QTL detection, comparative mapping, detection of chromosome duplications and marker-assisted selection. Genetic linkage maps have been constructed for maize using RFLPs (Helentjaris *et al*., 1986; Gardiner *et al*., 1993; Coe *et al*., 1995; Lee *et al*., 2002), AFLP (Vuylsteke *et al*., 1999), and SSRs (Sharopova *et al*., 2002). Also, sequenced cDNAs (also known as ESTs, expressed sequence tags) become a source of molecular markers and have now been integrated into maize genetic maps (Causse *et al*., 1996; Davis *et al*., 1999). These maps make it possible to locate genes and map QTLs in the maize genome.

2.6.1 Conventional method of QTL detection

QTL mapping is defined as association between observed trait values and the presence/absence of alleles of markers that have been mapped onto a linkage map. The established methods of detecting QTLs involves the selection of two parents that differ distinctly in a particular character and then determination of association between markers and that character in $F_{2,}$ and backcross progeny. If the correlation between the phenotype and alleles of the marker is significantly different from zero, then a QTL is detected.

In maize, a genetic map has been constructed from a large recombinant inbred line (RIL) population to increase the resolution of mapping (Lee *et al*., 2002). RILs are produced by inbreeding individual F_2 progeny up to six times to make them homozygous at any locus. Each RIL is thus fixed for short linkage blocks of parental alleles (Burr *et al*., 1988). RIL populations constitute permanent mapping populations

and can be used by different researchers in varying environments and the information can be added to a common database. As a result, the public maize-breeding sector has been able to develop detailed QTL and single gene maps for a number of traits (reviewed in Hoisington and Ribaut, 1998; Tuberosa *et al*., 2002). Despite these efforts, the resolution for many QTL maps is still several centimorgans (cM), corresponding to hundreds of genes.

Major shortcomings of QTL detection experiments include: (1) the limited number of recombination events per generation results in poor resolution, (2) only two alleles at any given locus can be studied simultaneously, (3) the number of location and effects of the identified QTLs vary according to the genetic background of the population, and (4) it is neither cheap nor fast (Tuberosa *et al*., 2002; Flint-Garcia *et al*., 2003). However, a number of alternative approaches of QTL detection are available whose application can contribute to partially circumvent some of the limitations discussed above.

2.6.2 Bulk segregant analysis

A cheaper and faster alternative to conventional QTL detection is bulk segregant analysis (BSA, Michelmore *et al*., 1991), which has been shown to work well with genes having major effects and that obviate the need for constructing detailed genetic map. For BSA of the trait of interest, parental lines are chosen that differ in their expression and crossed, and F_2 or RIL populations are generated which will segregate for the trait. The population is then phenotyped to identify individual plants or lines

having high or low expression of the trait. Two DNA bulks are prepared, one from the 'high' individuals and the other from 'low' individuals, and analyzed for allelic frequency with molecular markers. With dominant markers, only a few individuals are required in each bulk. The probability of an unlinked locus being polymorphic between two bulks of 10 individuals was calculated to be 2×10^{-6} (Michelmore *et al*., 1991). However, when using co-dominant markers (such as RFLPs and SSRs) with pools of genetically diverse individuals, where several marker alleles may be present, at least 50 individuals need to be combined to make each bulk (Quarrie *et al*., 1999)

The BSA can be used whether the individuals come from a single segregating population or from pools of genetically diverse individuals, such as variety mixtures or composite populations of out-breeding species such as maize. In maize, BSA has been efficiently used for the identification of QTLs for flowering time and yield (Tuberosa *et al*., 1998; Quarrie *et al*., 1999). However, a major shortcoming of BSA is that no information is provided on the distance of the QTL from the polymorphic marker, therefore, markers obtained in a BSA need to be mapped with standard approaches.

2.6.3 Association mapping

Another approach for QTL detection is association mapping. It is a population-based method used to identify marker-trait relationships based on linkage disequilibrium (LD, Remington *et al*., 2001). Linkage disequilibrium or allelic association is defined as the nonrandom association of alleles at different loci. Linkage refers to the

correlated inheritance of loci through the physical link on a chromosome, whereas LD refers to the correlation between alleles in a population.

Association and quantitative trait locus (QTL) studies suggested that the maize gene *Dwarf8* might affect the quantitative variation of maize flowering time (Thornsberry *et al*., 2001). In wheat this gene has contributed to yield increments seen in the 'Green Revolution' varieties and the *Arabidopsis* ortholog has been shown to play a role in regulating flowering time variation (Wilson *et al*., 1992). Similarly, association-mapping studies using RAPDs on genetically diverse rice germplasm (Virk *et al*., 1996) have identified markers associated with a number of characters, such as flowering time and panicle length.

The potential advantages of association mapping over conventional mapping are (i) only polymorphisms with extremely tight linkage to a locus with phenotypic effects are likely to be significantly associated with the trait in a randomly mating population, providing much finer resolution than genetic mapping (Remington *et al*., 2001), (ii) QTLs for any quantitative trait can be studied in the same investigation (Vuylsteke *et al*., 2000a), and (iii) detection of QTLs that vary across a wide spectrum of the gremplasm rather than just between two parental lines (Virk *et al*., 1996).

2.6.4 Comparative genetic mapping

One of the applications of genetic mapping is the comparison of genome colinearity and synteny within and between related crop species. Colinearity is defined as the conservation of gene content and order between two or more species, while synteny is

defined as the conservation of linkage on chromosomes, in the absence of a defined order (Bennetzen and Devos, 2002). Because of their conserved genetic nature, some DNA markers can be used in genetic mapping of the species of origin and closely related species. For example, species of the Poaceae (maize, sorghum, rice, oat and wheat) share conserved gene collections. Of 150 maize RFLP markers tested, only one failed to hybridize to sorghum DNA (Hulbert *et al*., 1990). About 85% of rice, oat and barely cDNA clones showed hybridization to maize DNA (Ahn and Tanksley, 1993). Hence, the same set of RFLP probes derived from a single species can be used for genetic mapping in related species. Thus, it is possible to compare linkage maps and determine whether the order of markers along the linkage groups is conserved across species.

However, detailed comparisons of genome colinearity and synteny can only be accomplished by comparative physical mapping and sequencing. This will provide new insights into gene and genome evolution, and are powerful tools for gene isolation and characterization. One approach of QTL cloning in maize is based on the identification and mapping of a large number of ESTs whose mapping will provide candidate genes for the QTLs (Davis *et al*., 1999). The maize genome will be sequenced providing the ultimate resources of candidate genes for QTL mapping and cloning.

2.7 Marker-assisted selection and breeding

Conventional plant breeding is time consuming and very dependent on environmental conditions. Breeding a new variety takes eight to twelve years and even then the

release of an improved variety may not be granted. Hence, breeders are interested in new technologies that could make this procedure more efficient. When selection is based on genetic information through the application of molecular markers it is called marker-assisted selection (MAS). Marker-assisted selection is based on the concept that it is possible to infer the presence of an allele of a gene from the presence of a marker allele that is tightly linked to the gene.

MAS improves selection for quantitative traits because (1) DNA markers can be assayed at the seedling stage, permitting one to make selections before many traits are expressed, thus reducing the number of individuals which must be grown to maturity, (2) many traits may be more accurately selected for by using genotypes at DNA markers than by relaying solely on phenotype which may be due to either genotype or environment, and (3) unlike phenotypic traits, genetic markers can be reliably assayed in non-target environments such as the growth chamber or greenhouse, permitting more rapid progress in breeding.

In principle, once QTLs have been identified, introgression of the favorable alleles and their pyramiding into elite germplasm (e.g. parental lines, populations, etc.) becomes possible through MAS (Ribaut and Hoisington, 1998, Stuber *et al*., 1999). However, only a few successful applications of MAS for improvement of quantitative traits have been described (Ragot *et al*., 2000: Ribaut *et al*., 2000; Bouchez *et al*., 2002) due to mainly to weak associations (in terms of genetic distance) between markers and target QTLs and high cost of MAS (Stuber *et al*., 1999; Moreau *et al*., 2000, Tuberosa *et al*., 2002).

2.7.1 Introgression of desirable genes

Another application of marker-assisted selection is the introgression of desirable genes from wild species into an elite variety. Tanksley and McCough (1997) proposed that wild or unimproved accessions may harbor important genes that can significantly improve yield and other important traits when introgressed into adapted cultivars with the use of DNA markers. In conventional plant breeding, backcross breeding is a well-known method for the introgression of desirable genes from a donor lines into recipient lines. Such components can be transferred to elite cultivated materials by repeated backcrossing. One of the disadvantages of this method is that other genes may also be transferred along with the genes that control the target trait, which may reduce yield or quality of the desired varieties. By the use of markers linked to specific QTLs it is possible to introgress specific regions of the genome that confer desirable quantitative characteristics to an elite variety (Tanksley *et al*., 1996; Harjes *et al*., 1999).

Given the results already produced in maize at the molecular level: linkage map (Davis *et al*., 1999; Lee *et al*., 2002), and QTL analysis (Veldboom *et al.,* 1994; Ribaut and Hoisington, 1998; Tuberosa *et al*., 1998), marker- assisted selection for maize improvement is becoming more and more efficient (Stuber *et al*., 1999; Bouchez *et al*., 2002; Tuberosa *et al.,* 2002).

2.8 Conclusions

Plant breeding relies on genetic variation and uses selection to improve plant productivity. Over 50% of agricultural productivity in the world has been achieved through traditional plant breeding (Kumar, 1999). However, as the human population increases and the expected reduction of available arable land due to climate and human intervention continues, it may be necessary to accelerate the rate at which genetic improvement is achieved.

Modern biotechnology provides new tools that can facilitate the development of improved plant breeding methods and expand our knowledge of plant genetics. The knowledge that is obtained with these new tools can be used to enhance food security throughout the world. Particularly, DNA markers have the potential to enhance the operation of plant breeding programs ranging from fingerprinting of genetic stocks, assessment of genetic diversity, increasing the efficiency of selection, to comparative mapping and manipulation of QTLs. Despite this potential, the current application of crop biotechnology is almost nil in most African countries, because of the lack of resources, trained personnel and infrastructure in this field.

The primary resource of plant breeding programs in Africa is the genetic variability available within landraces or primitive varieties. The success of crop improvement is highly dependent on the power and efficiency with which this genetic variability can be manipulated. However, in many African countries plant breeders still use morphological traits to study genetic diversity and genetic relationships among genotypes.

Morphological differences do not always reflect genetic differences, because of genotype x environment interaction. As a result, the potential of making genetic progress is slow. Therefore, in the future it is necessary to use DNA markers that will provide more rapid and precise information on the extent of genetic diversity and genetic relationships among genotypes than phenotypic selection.

Recent advances in automated marker technology have presented the possibility of efficiently applying marker-assisted selection at the scale of modern plant breeding. Conventional QTL studies are commonly based on the phenotypic and molecular analysis of single genotypes (individual plants or progenies) of a mapping population mostly derived from the cross of inbred lines. However, cheaper and faster alternatives to conventional QTL detection (e.g. BSA and association mapping) can be effectively used in African crops, which do not require the development of RILs and mapping populations.

Large-scale explorations of plant genomes will rapidly narrow the gap in knowledge between the model crops (rice and maize) and lesser-studied African crops (e.g. sorghum, millet and teff). Based on the intensive study of genes for important agronomic characters in rice or maize, it may be possible to make rapid developments of these traits in sorghum and millet breeding. Therefore, simple PCR-based markers (SSR, RAPD and AFLP) are an appropriate entry point to genomics for many African countries, including Ethiopia.

In the highland areas of Ethiopia, maize is the most important crop grown by

subsistence farmers and it is an important "hunger breaking crop" due to the fact that it is often consumed green. It is hypothesized that many of the highland maize varieties have been geographically isolated for long periods of time and may have accumulated specific genetic adaptations for highland conditions. Therefore, it is necessary to study the genetic diversity and genetic relationships among these accessions using morphological and molecular markers in order to (a) understand the distribution of genetic variation in different highland regions, (b) better conserve the genetic variation contained in them and (c) facilitate their use in new, dedicated breeding programs for highland maize.