Journal of Agricultural Science; Vol. 15, No. 3; 2023
ISSN 1916-9752 E-ISSN 1916-9760
Published by Canadian Center of Science and Education

Molecular Markers in Plant Breeding

Luciana L. Benchimol-Reis[1]

[1] Centro de Pesquisa em Recursos Genéticos Vegetais, Instituto Agronômico-IAC, Campinas, SP, Brazil

Correspondence: Luciana L. Benchimol-Reis, Centro de Pesquisa em Recursos Genéticos Vegetais, Instituto Agronômico-IAC, Campinas, SP, Brazil. Tel: 55-19-3202-1801. E-mail: luciana.reis@sp.gov.br

Received: January 11, 2023 Accepted: February 3, 2023 Online Published: February 15, 2023

doi:10.5539/jas.v15n3p58 URL: https://doi.org/10.5539/jas.v15n3p58

Abstract

Molecular markers are an important tool for plant breeding. Since the 1980s, in response to the technology development, molecular marker approaches have been further diversified. The establishment of new-generation sequencing and high-throughput plant phenotyping has greatly decreased the time to genotype large numbers of individuals. For breeders who are not very familiar with molecular techniques and want to catch up with the advances in the field, this review offers basic knowledge. Each molecular marker technology has specific advantages as well as limitations. Molecular marker types, diversity studies, QTL mapping, associative mapping, marker-assisted backcrossing and genomic selection are explored. Marker application in plant breeding is also described. In the genome, molecular markers can detect the genetic architecture of a trait, but also identify candidate genes with an important role in plant breeding programs.

Keywords: SSRs, SNPs, DARTs, GBS, GWAS, GWS

1. Molecular Markers

A marker is any morphological or molecular characteristic that allows the differentiation of individuals. Figuratively speaking, in a puzzle (genome of a species) where each piece is “marked” with a number, it is much easier to locate each one and its neighbors (tagged genome, Figure 1).

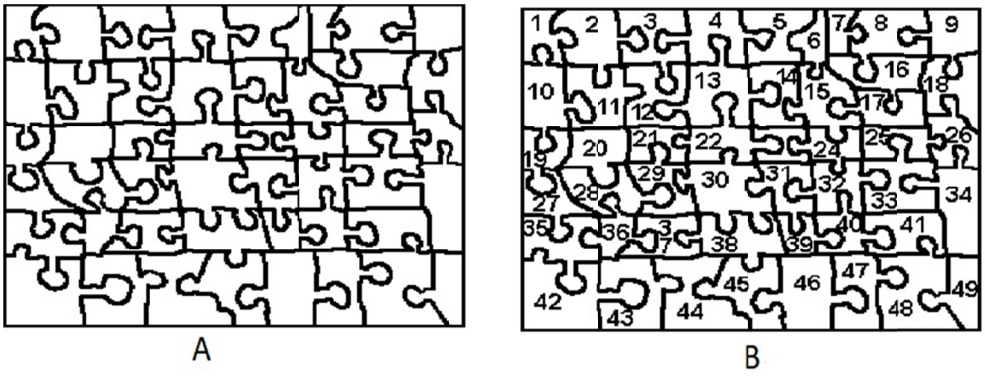


Figure 1. Genome (A) without markers and (B) with molecular markers

Markers could also be compared to “flags” in the genome that make it simpler to locate genes close to or exactly where these flags stand (Collard, Jahufer, Brouwer, & Pang, 2005). In the beginning, morphological or phenotypic markers were used (Karaköy, Baloch, Toklu, & Özkan, 2014), but their polymorphism, *i.e.*, variability (Figure 2), was very low, because they depend on trait expression, where diversity is lower.

(A)

(B)

Figure 2. A) Monomorphism, *i.e.*, absence of variability; B) Polymorphism, *i.e.*, presence of variability in maize grain color

In the 1960s, the biochemical markers were introduced, of which isoenzymes are particularly relevant. These are enzymes that differ in the amino acid sequence but catalyze the same chemical reaction (Penteado, García, & de la Veja, 1997). As they depend on the variability observed because of RNA-to-protein translation, polymorphism was also low.

In the 1980s, the first molecular marker at the DNA level was proposed. Molecular markers are independent of the phenotype, detect variations at the DNA level, are not affected by environmental influence, have high heterozygosity, a large number of available markers and Mendelian inheritance. An important concept to differentiate molecular markers is their dominance or codominance (Figure 3).

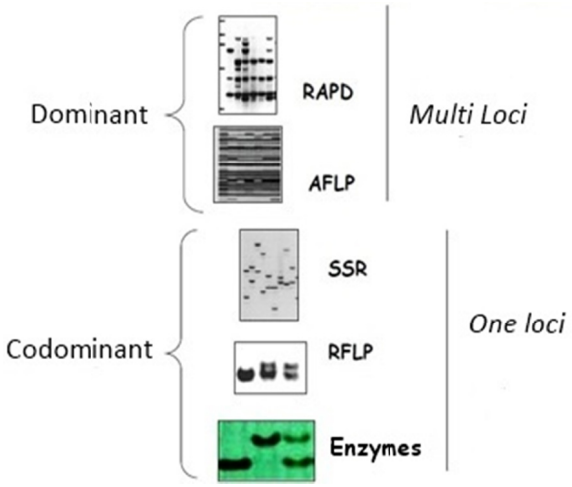


Figure 3. Dominant and codominant molecular markers. Random Amplified Polymorphism (RAPD), Amplified Fragment Length Polymorphism (AFLP), Simple sequence repeats (SSRs) and Restriction Fragment Length Polymorphism (RFLP)

A dominant marker cannot differentiate a homozygous from a heterozygous individual (Figure 4). In contrast, the heterozygote of a codominant marker can be differentiated (Figure 5). Mendelian inheritance is not an absolute requirement, because sometimes a molecular marker can fall within a transposon that does not have Mendelian inheritance, and nevertheless, the marker can be functional (Figure 6).

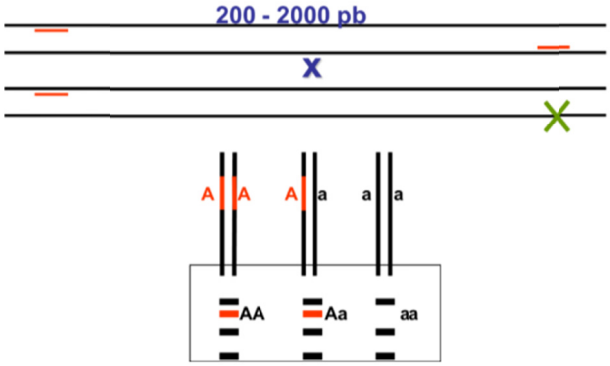


Figure 4. Example of a dominant marker

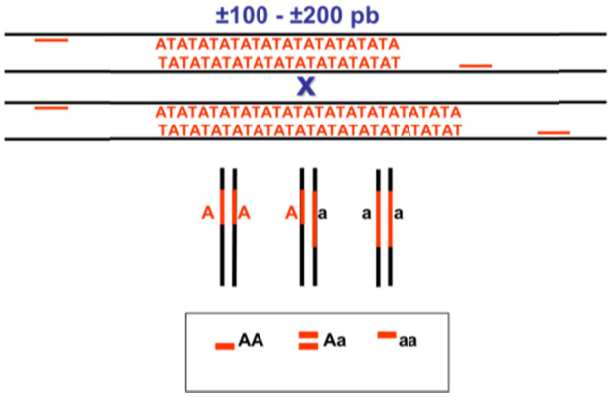


Figure 5. Example of a codominant marker

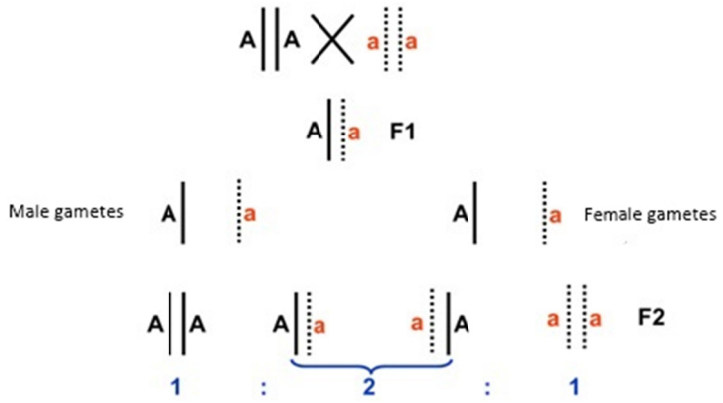


Figure 6. Mendelian segregation of a molecular marker

The first molecular marker was Restriction Fragment Length Polymorphism (RFLP, Botstein, White, Skolnick, & Davis, 1980), which was first developed in human genetics and later applied to plants. This technique examines the difference in size of restriction fragments of DNA cut by restriction enzymes. Once cut, the fragmented DNA is run on a gel (electrophoresis) and then analyzed by Southern Blotting, a procedure by which the DNA is transferred by capillarity from the gel to a nylon or nitrocellulose membrane. Once loaded on the membrane, this membrane is hybridized with a probe (of genomic DNA or cDNA—complementary DNA) in a temperature-controlled hybridization chamber. Fragments homologous to this probe will hybridize to the membrane. The probe is labeled prior to hybridization with radioisotopes or digoxigenin (cold labeling). Polymorphism results from point mutation, fragment insertion or deletion, restriction site translocation along with probe hybridization. Because RFLP uses restriction enzymes, pure high-molecular weight DNA is required. The RFLP *locus* is the probe combination with the restriction enzyme used (Figures 7, 8 and 9).

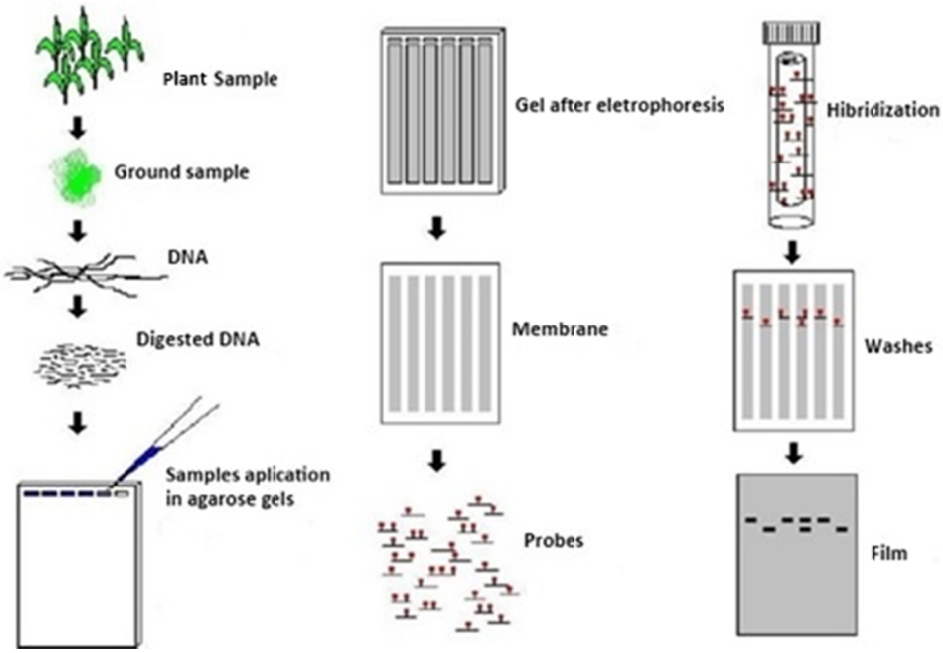


Figure 7. Scheme of the RFLP technique (Sibov et al., 2003). Total genomic DNA is digested by a restriction enzyme which forms a smear on the stained gel; the DNA is transferred to a membrane by Southern blotting; hybridization of probes to the DNA on the membrane takes place.
Only bands hybridized to the probes were taken into consideration

The advantages of this technique are that whole-genome coverage is possible (variations in DNA sequences from 4 to 8 bp are detectable) and that it is codominant and has high repeatability and consistency. The limitations are intensive steps in labor, lack of probe library available and laboratory structure.

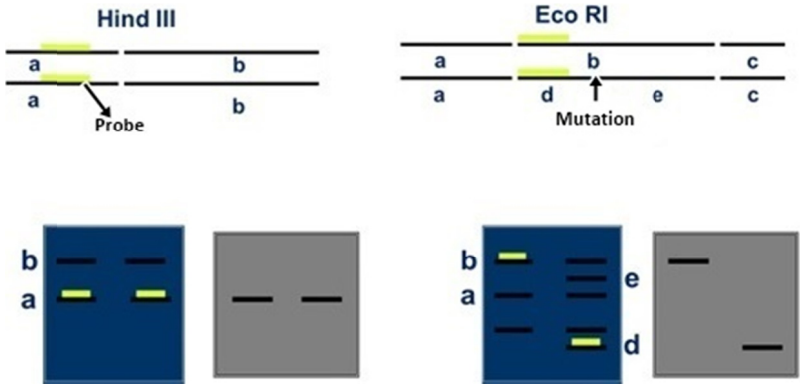


Figure 8. RFLP analysis with the same probe and different enzymes generates differences in the observed fragments

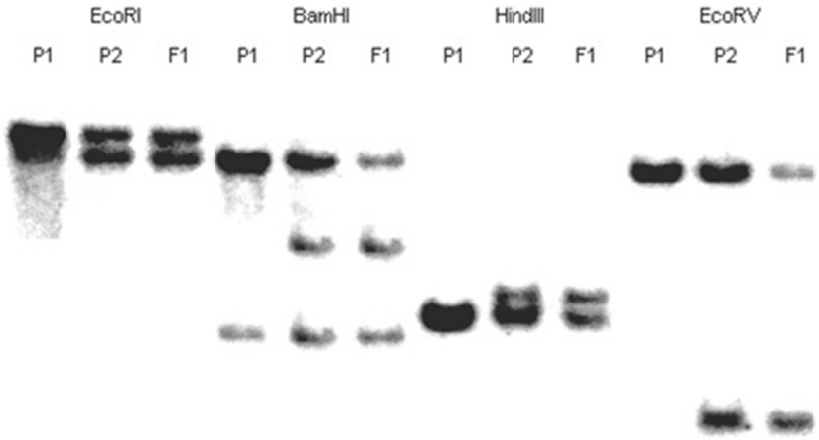


Figure 9. RFLP from the same maize DNAs (P1: parental DNA 1; P2: Parental DNA 2; F1: First filial generation, heterozygote) with the same probe (Umc1546) and different restriction enzymes (*Eco*RI, *Bam*HI, *Hind*III and *Eco*RV). Sibov et al. (2003). Each probe + enzyme combination is a *locus*

In 1990, after the discovery of the Polymerase Chain Reaction (PCR, Mullis, & Faloona, 1987) and the Random Amplified Polymorphic DNA (RAPD; Williams, Kubelik, Livak, Rafalki, & Tingey, 1990), the marker system was launched. With only a single short random primer, the system amplifies anonymous DNA sequences (10 bases with > 50% G + C). The primer hybridizes around in different parts of the genome. The method can quickly detect polymorphisms, but is not very informative, for being dominant. Furthermore, the poor reproducibility and co-migration are drawbacks: it is not known whether the same band is the same fragment of the genome. Results can only be proven by band elution and sequencing. The gel is read by taking the molecular weight of each bandinto consideration (Figures 10 and 11).

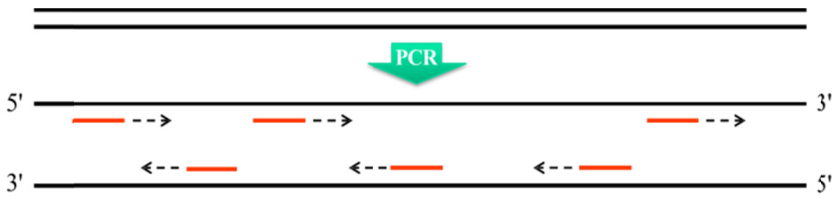


Figure 10. Representation of RAPD primer annealing a in homologous parts of the genome

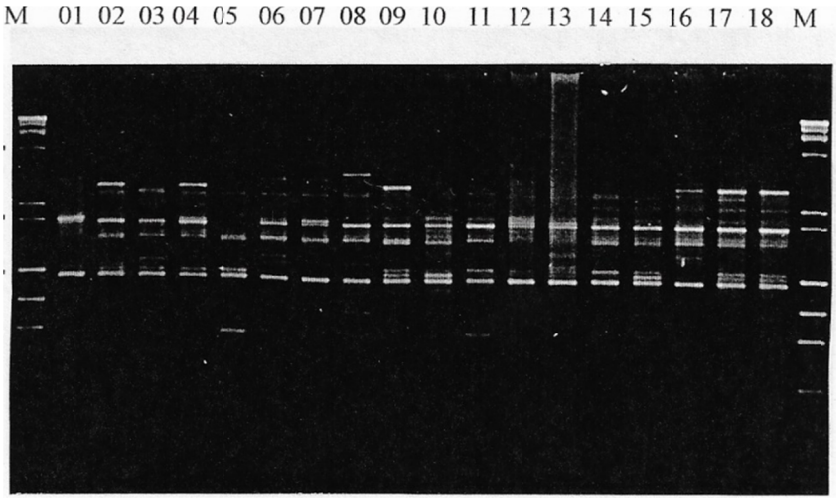


Figure 11. RAPD in 18 inbred maize lines with OPB-09 (TGGGGGACTC). M is the molecular marker in pb

In 1993, the Sequence-Characterized Amplified Region (SCAR) marker (Paran & Michelmore, 1993) was launched. This single-locus marker is derived from eluted, cloned and sequenced RAPD or Amplified Fragment Length Polymorphism (AFLP) fragments. These markers are more stable due to specific primers, but are still dominant and analyze band presence/absence.

In 1995, the technique of Amplified Fragment Length Polymorphism (AFLP) (Vos et al., 1995) was announced. It first cleaves genomic DNA with two (a frequent cutting and a rare cutting) restriction enzymes and ligates specific adapters (20 - 30 bp) to the ends of the cleaved fragments. Ligated DNA samples are amplified by PCR. In a first step, the so-called pre-amplification, one in every 04 bases (A, T, C, G) is edited, generating 1:16 amplified fragments (Figure 12). Thereafter, 1: 4096 fragments are generated by selective amplification (primer with three selective bases). In this step, a rare-cut restriction enzyme is used (Figure 13). Amplified Fragment Length Polymorphism is dominant, in other words, it does not distinguish heterozygotes.

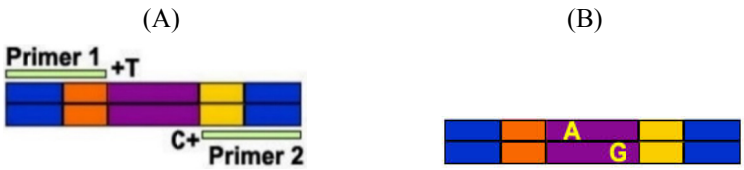


Figure 12. A) and B) Selective pre-amplification generating a subpopulation of complementary fragments with complementary bases to those used for the selective primers

Of the cleaved fragments, large fragments obtained from rare-enzyme cutting are concentrated at both extremities (4-5 bp fragments). Intermediate fragments generated by rare and frequent cuttings with optimal resolution are found in the middle of the gel, while small fragments generated by frequent enzyme cuttings are observed at the bottom of the gel (Figure 14).

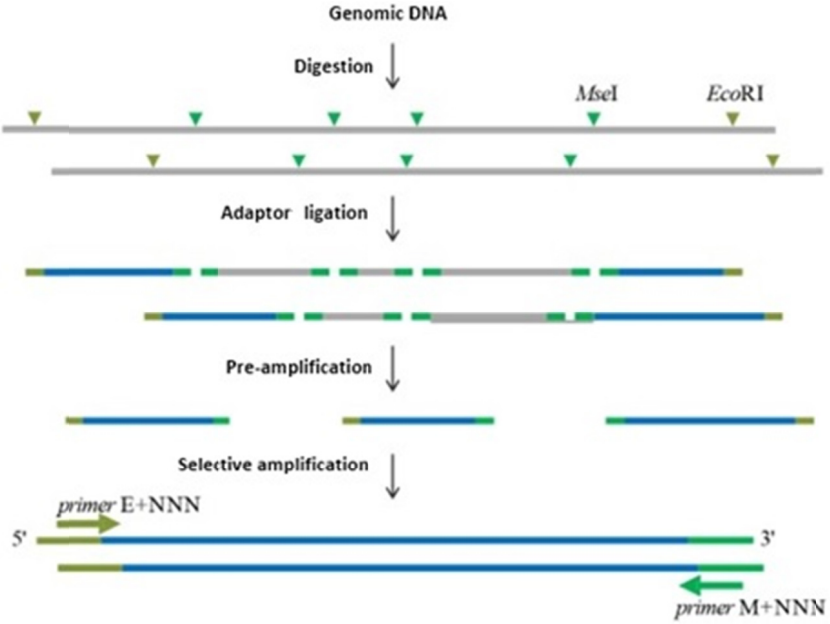


Figure 13. Schematic representation of the AFLP steps

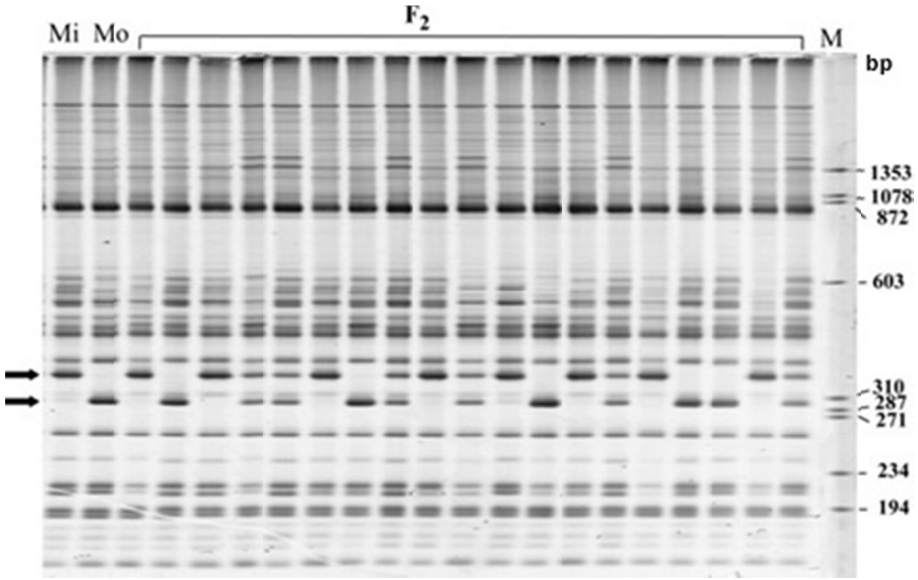


Figure 14. AFLP gel for soybean parental genotypes (Mi: Misuzudaizu and Mo: Moshidou) and the F2 population. Xia et al. (2007)

Microsatellites or Simple Sequence Repeats (SSRs) are short sequences (1-6 bases) repeated in "tandem" (Hamada, Petrino, & Kakunaga, 1982; Tautz & Renz, 1984), widely found in the genome of different species (from prokaryotes to eukaryotes). They are present in coding and non-coding regions. They also occur in organelles such as chloroplasts and mitochondria. Owing to natural mechanisms of mutation, the variability is considerable (Oliveira, Pádua, Zucchi, Vencovsky, & Vieira, 2006), resulting in variation in the number of copies of the basic repeat unit (Schlötterer & Tautz, 1992). The most common motif in plants is that of dinucleotides, especially of AT repeats (Kalia, Rai, Kalia, Singh, & Dhawan, 2011). Microsatellites are codominant markers and quite informative, as heterozygous individuals can be detected. Although the regions of the microsatellites are highly variable, the flanking sequences are conserved, on which the primer pairs for marker amplification are designed (a forward and a reverse primer, Figures 15 and 16).

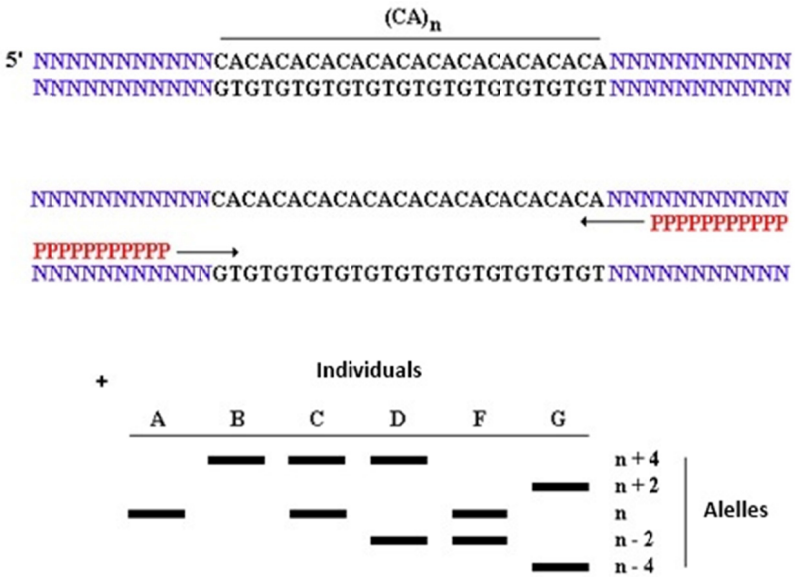


Figure 15. Microsatellite markers (CA)n. Each variation in the motif generates a different allele at the same *locus*

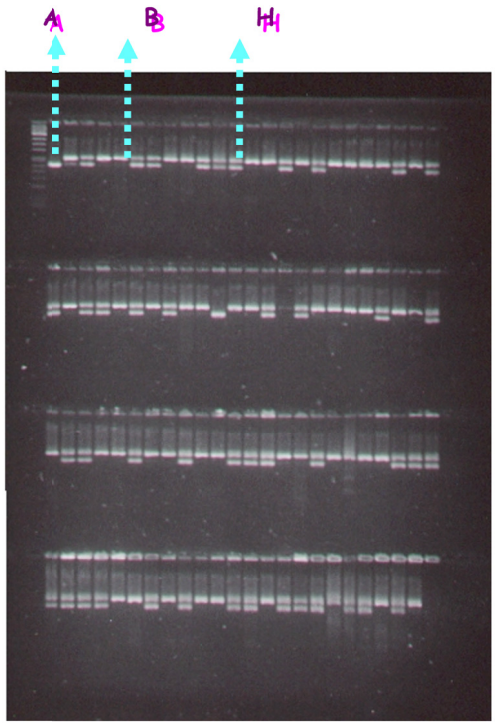


Figure 16. A and B are genotypes of maize parents of the F1 hybrid (H). Microsatellite PHI037 was amplified and run on a 4% MetaPhor agarose gel, stained with ethidium bromide

The microsatellites can be derived from a database, by identifying microsatellite-containing sequences. Alternatively, microsatellite-enriched libraries can be produced. In these, genomic DNA is completely fragmented by a restriction enzyme (example *Rsa*I). Then, adapters are ligated to these DNA sequences, followed by hybridization with repeat-rich primers such as the biotin containing $(GT)_n$. These primers will hybridize in genome regions where there is a complementary repeat, in this case, $(CA)_n$. Next, the mixture is exposed to streptavidin which has an affinity for biotin. The fragments are separated with streptavidin-coated magnetic beads (Streptavidin MagneSphere® Paramagnetic Particles, Promega) so that the fragments with $(CA)_n$

repeats are separated from the rest of the sample. Next, PCR amplification is performed with primers complementary to the adapters. These amplified fragments are cloned into a plasmid and *Escherichia coli* are transformed with plasmids (pGEM-T vector) that carry enriched sequences. These will be sequenced for microsatellite confirmation and to design the primers on conserved flanking sequences.

Microsatellites have codominant inheritance, high abundance in eukaryotic genomes, multi-allelicity, low cost and potential for amplification of multiplex systems, while execution is easy, transferable and automatable. The development of specific primer pairs through the construction of microsatellite-enriched genomic libraries may be less advantageous, as the library must be cloned and sequenced (Figure 17), which can be costly and laborious.

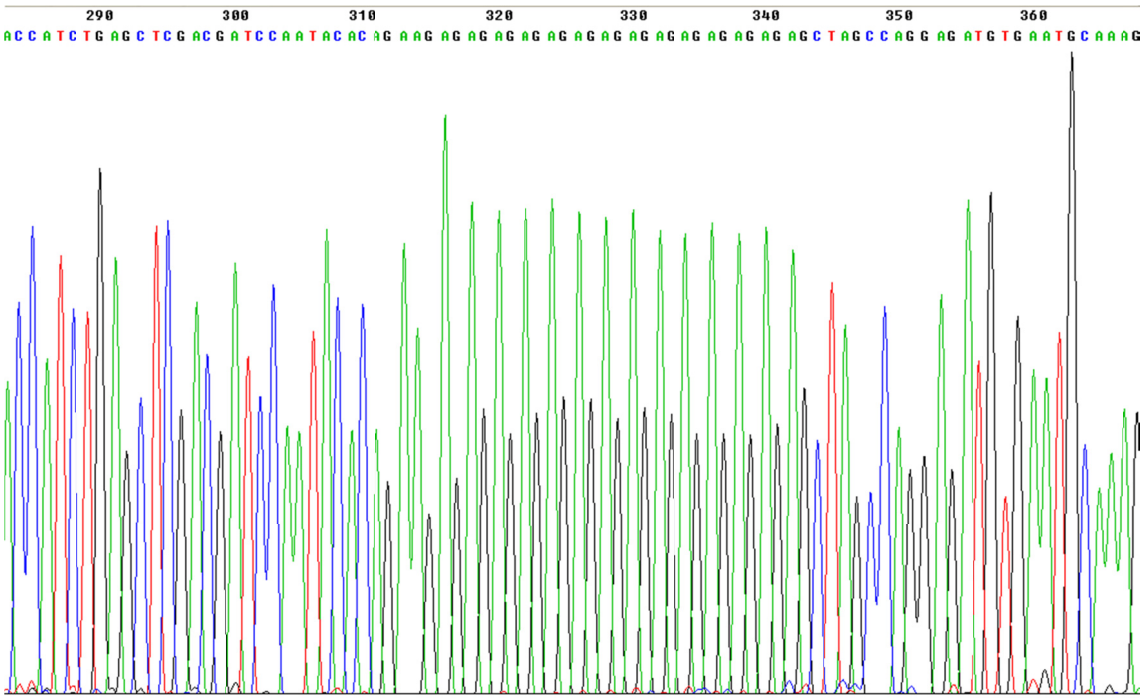


Figure 17. Electropherogram of a repeat of a microsatellite $(AG)_{14}$

The Inter-Simple Sequence Repeat (ISSR) is a dominant marker with low development cost, a high polymorphism rate and high reproducibility (Chen et al., 2017). Due to the lower polymorphism, such markers are less effective than AFLP (Costa et al., 2016). They have the advantage that they can be designed without previous sequence knowledge and can be detected on agarose or polyacrylamide gels.

Target Region Amplification Polymorphism (TRAP) is a PCR-based technique. It uses candidate genes (ESTs-Expressed sequence Tags) and increases the possibility of the genetic variability (Hu & Vick, 2003). It employs two 18-base primers: the forward is a fixed primer, designed from an EST sequence of interest. Its characteristic is to have 18 nucleotides and an optimal, minimum and maximum Tm (melting temperature) set at 53, 50 and 55 ^{o}C, respectively. The reverse primer is a arbitrary, with the following characteristics: selective nucleotides at the 3′ end, 4-6 AT or GC nucleotides in the central region, 10-11 nucleotides as filler sequences at the 5′ end. The polymorphism of TRAP is given by having a fixed primer with several arbitrary primer combinations. Visualization on a DNA sequencer is by fluorescent labeling of arbitrary primers or by silver staining, resulting in an AFLP-like band pattern.

Single-nucleotide polymorphism (SNPs, Hwang et al., 2016) or single-base polymorphism occurs in more than 1% of the population. SNPs are mutations propagated over generations. They are abundant in the genome, codominant and biallelic. They can be by transversion (purine to pyrimidine)—A by C or T, G by C or T—or by transition (purine by purine; pyrimidine by pyrimidine)—A by G, C by T (Figure 18). They can occur in expressed and non-expressed regions; they are stable from an evolutionary point of view and genome frequency and distribution is high. In the human genome, the average distribution is 1 SNP every 300 bp of sequence

(Nelson et al., 2004); in maize, there is 1 SNP every 48 bp (non-coding regions, Batley et al., 2003), with a total of 21,502 SNPs, and 1 SNP per 3.3 kb and a total of 25,274 SNPs in Arabidopsis (Drenkard et al., 2000).

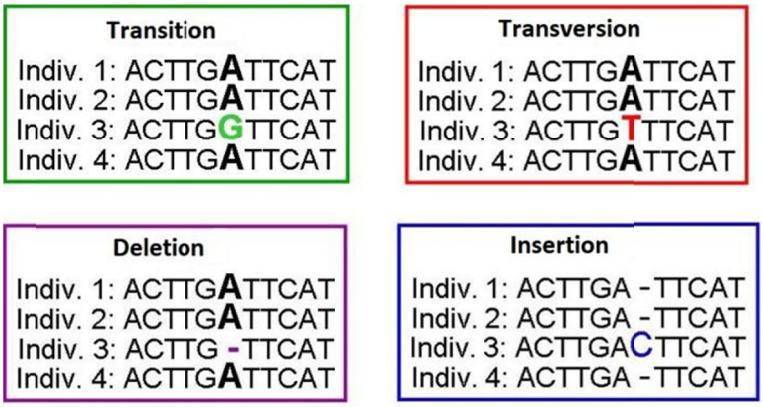


Figure 18. Different kinds of base alteration in SNPs

The SNPs can be identified in two ways: *in silico*—by data analysis of sequences available in databases; *in vitro*—a new data sequence is generated. The *in silico* approach uses software to identify SNPs in a sequence bank, since manually it would be impossible to identify SNPs in many sequences. Several programs are available such as Phred, PolyBayes, SNPserver, etc. The advantages of using SNPs are the abundance in the genome, detection of different alleles for genes of interest (bi-allelic), high accuracy, high reproducibility and a high level of automation. The rise in use of SNPs follows the emergence and evolution of next generation genotyping platforms (NGS).

The SNP variants form blocks called haplotypes (Figures 19 and 20). Haplotypes are sequences in a cluster that represent the same allele of a gene. In this case, two or more SNPs segregate together (Rafalsky, 2002). In other words, they have the same nucleotide at the polymorphic site. Mathematical methods are used to minimize false reconstruction of haplotypes caused by sequencing errors.

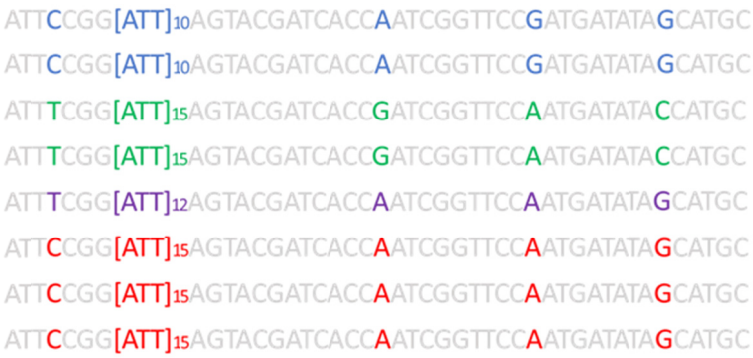


Figure 19. Example of haplotype with SNPs in linkage disequilibrium (LD)

In this way, together these SNPs will be characterizing a linkage disequilibrium (LD), which is simply the lack of independent segregation between the alleles at two or more *loci*. In general, LD depends on the history of recombination between polymorphisms. Factors such as genetic drift, selection between populations, migration (gene flow) and reduction in population size ("bottleneck") can modify the LD between markers and related traits (Figures 21). The SNP tag is the smallest set of haplotypes that can differentiate individuals (Figure 22).

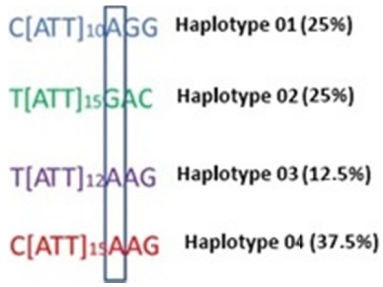


Figure 20. Example of different haplotypes. Assume that the G allele increases the growth rate

Figure 21. The C allele in the last position serves as a tag for the G allele. We say that these two positions (of the G and C allele) are in linkage disequilibrium (LD)

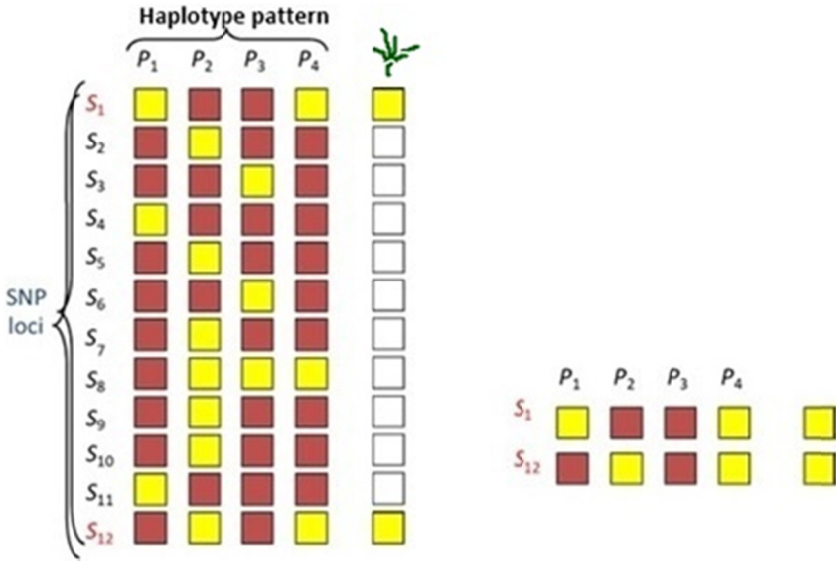


Figure 22. S1 and S12 SNPs can form a set of SNP tags. This is the smallest set of SNP tag in this example

A new concept called genotyping-by-sequencing (GBS) was conceived (Elshire et al., 2011) at the Buckler lab, Cornell University, where the detection of sequence differences (SNPs) in a large segregating population was combined with scoring thus allowing a quick and direct study of its target diversity to the mapping of a trait of interest. Lower costs in NGS have made GBS accessible for application in different plant species. The popularity of SNPs is still on the rise with NGS platforms. Be it by GBS or Beadchips (Illumina), SNPs are nowadays the most popular markers in plant breeding.

The Diversity Arrays Technology (DArT; Jaccoud, Peng, Feinstein, & Kilian, 2001) is based on the comparison of genome representations, does not require previous sequences (genome), nor depend on gels, can be developed for genomes of any size and is relatively inexpensive. The technology is based on methods of complexity reduction by Restriction Enzymes (ER), digestion, adapters, PCR (Figures 23), construction of a Genomic

Library and then microarrays. To this end, the technology requires the development of diversity panels for a given species, fluorescence marking, hybridization and image analysis (Figure 24). It must be mentioned that the DART technology is in the hands of a single company in Australia (DART PTy, https://www.diversityarrays.com/). These markers can be used for: germplasm characterization, marker-assisted selection (MAS), identification of target genes, construction of genetic maps, QTL mapping (Quantitative Trait *Loci*) and GWAS (genome-wide association studies).

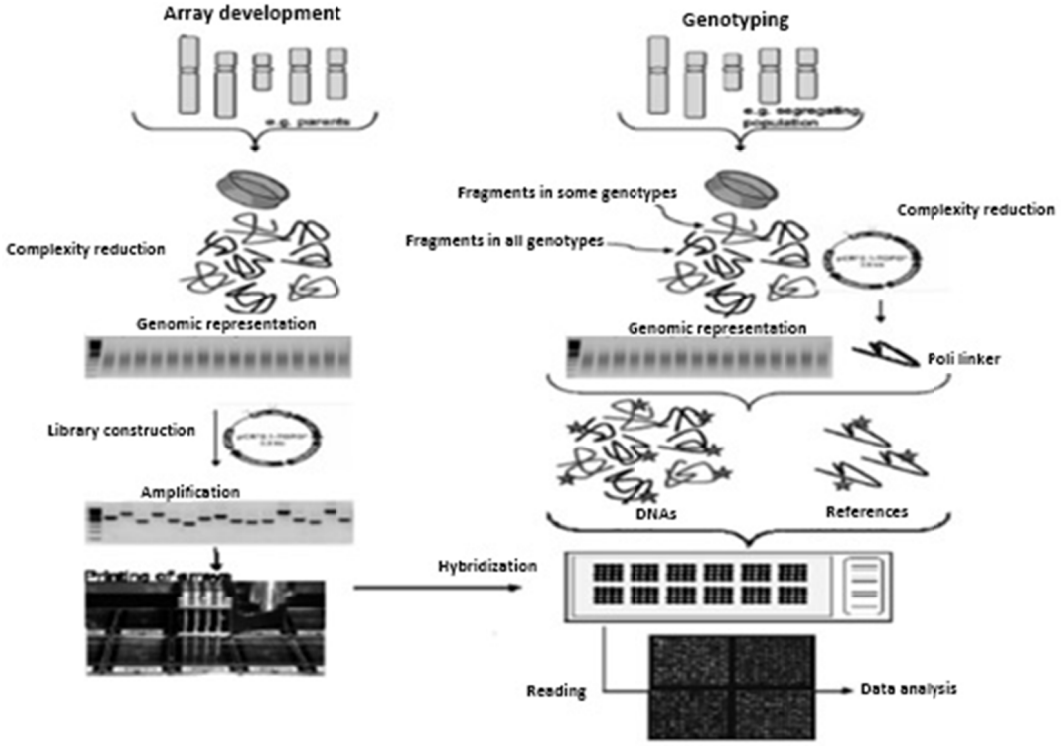


Figure 23. Development stages of diversity panels and DART marker genotyping

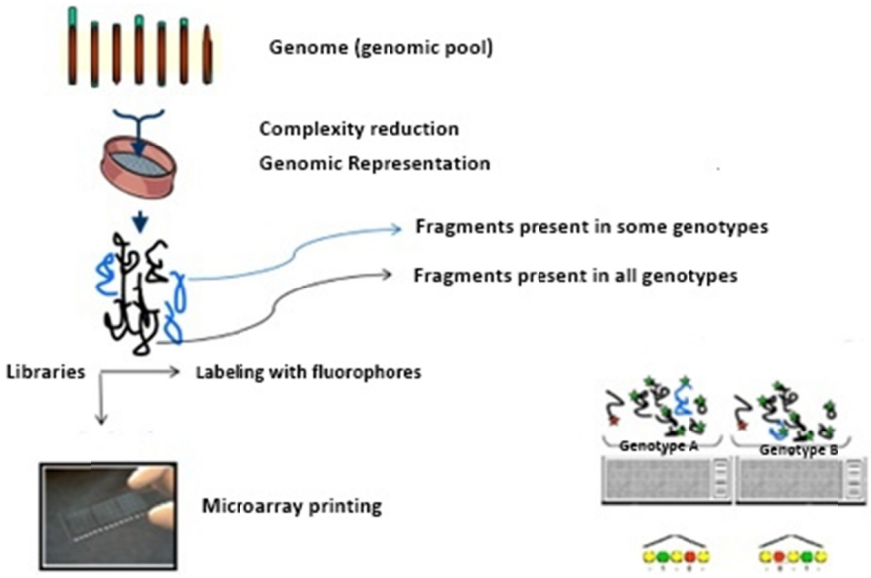


Figure 24. Genotyping with DART markers

2. Application of Molecular Markers

2.1 Genetic Diversity Analysis using Molecular Markers

After genotyping, the first step is to elaborate the data matrix where each band absence is read as "0" and each presence as "1". In the case of dominant markers, a genetic similarity coefficient is applied to the matrix, e.g., the Jaccard coefficient that excludes the double absence in its formula, resulting in a triangular similarity matrix. When working with distant species it is better to use the Dice coefficient, which assigns double weight to similarities (a) between species. These similarity coefficients can be calculated by the NTSYS-PC program, version 2.0j (Rolf, 1997). In the case of codominant markers such as microsatellites, it is used to apply a genetic distance coefficient such as Rogers Modified Distance (Goodman & Stuber, 1983). The latter can be calculated using the TFPGA program, version 1.3 (Miller, 1997). After establishing the similarity (S) or distance (D; D = S − 1) matrix, a clustering coefficient is applied to obtain a spatial representation of individuals called a dendrogram (Figure 25). One can also use the distance and/or similarity matrix to perform multivariate analysis, e.g., Principal Coordinate Analysis (Figure 26).

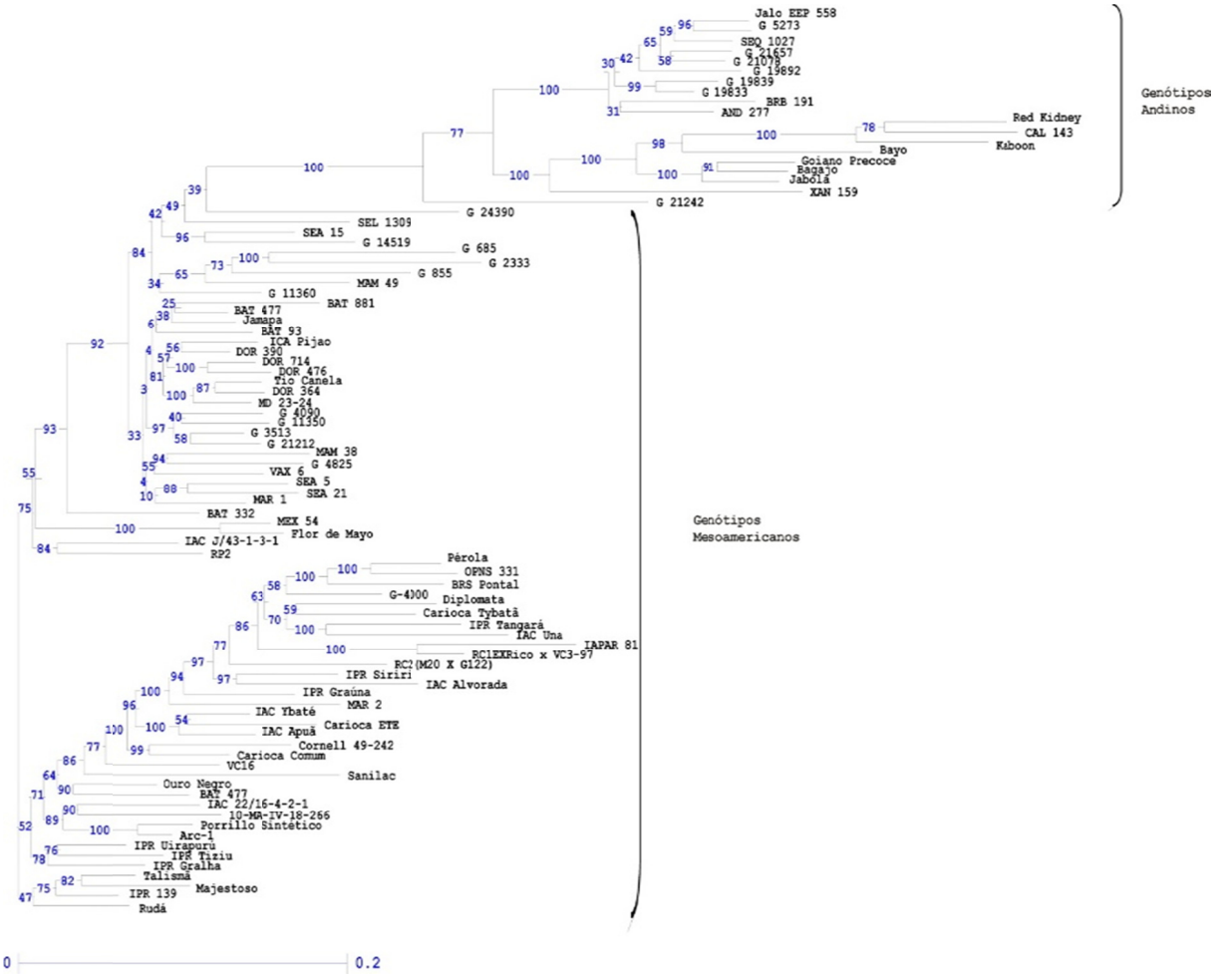


Figure 25. Analysis of diversity in common bean with 2,501 DART markers (Briñez et al., 2012; Reproducibility: 99.9%)

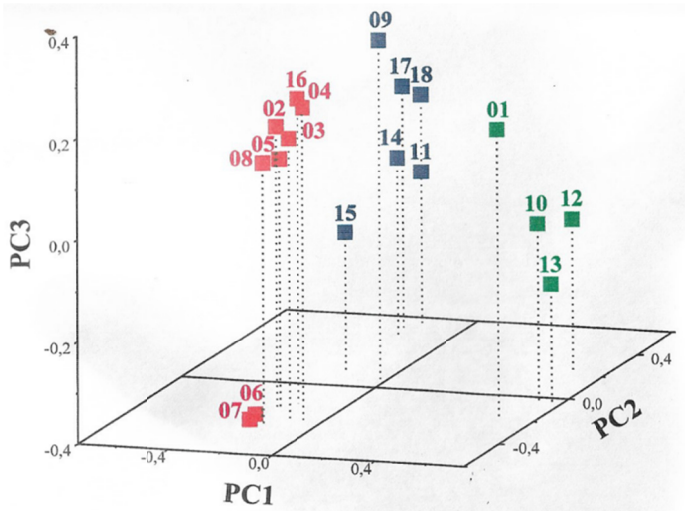


Figure 26. Principal Coordinate Analysis of 18 maize inbred lines with 262 polymorphic RAPD bands

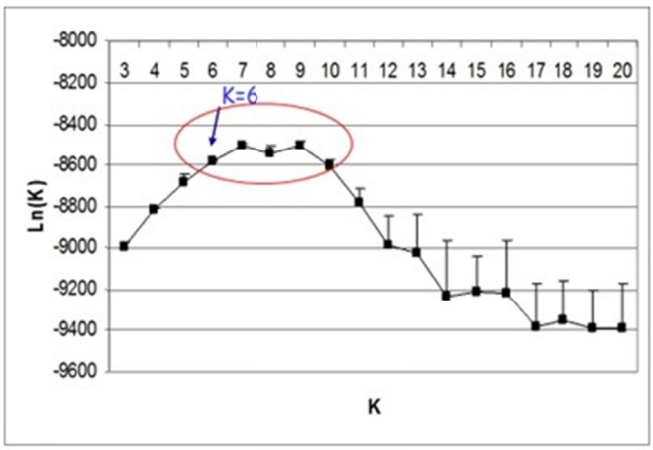
To calculate the informativeness of each molecular marker, the following formula for the polymorphic information content (PIC, Equantion 1) is used:

$$PIC = 1 - \sum_{i=1}^{n} f_i^2 - \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} 2 f_i^2 f_j^2 \quad (1)$$

Where: n is the number of alleles and f_i and f_j are the frequencies of the i^{th} and j^{th} alleles, respectively (Botstein et al., 1980).

The sample variance of the genetic distances is also calculated by Bootstrap analysis, with at least 500 re-samplings with repetitions (Tivang, Nienhuis, & Smith, 1994). By this analysis, each branch of the dendrogram is statistically represented in percentage.

In addition to these analyses, Bayesian statistics that use likelihood can be included. The STRUCTURE program (Pritchard et al., 2000) generates clusters based on transient Hardy-Weinberg disequilibrium (HW) and LD caused by genetic mixing between populations. The appropriate first step is to calculate the likelihood of the data for a range of K values (number of groups) by creating posterior probabilities of K, called X and written: X|K. For each dataset, the first step is to identify the most likely K. Two methodologies can be used for this purpose: Pritchard and Wen (2004) (Figure 27) and Evano, Regnaut, and Goudet (2005) (Figure 28). After determining the best K, the genotypes are assigned to the groups (Figure 29).



Number of groups	Values Ln(k)	Standard deviation
3	-8993.9	26.79
4	-8813.91	2.53
5	-8685.97	45.12
6	-8581,16	1,61
7	-8508.58	9.56
8	-8541.84	37.48
9	-8507.86	23.11
10	-8606.61	34.64
11	-8783.93	71.24
12	-8988.62	143.23
13	-9027.51	191.64
14	-9233.95	269.49
15	-9212.74	170.38
16	-9215.67	254.26
17	-9384.12	208.33
18	-9351.04	192.82
19	-9387.55	184.48
20	-9393.33	223.86

Figure 27. Graphical representation of the ideal number of groups (K = 6) inferred from the criterion of Pritchard and Wens (2004) for 60 rubber trees analyzed with 68 SSRs

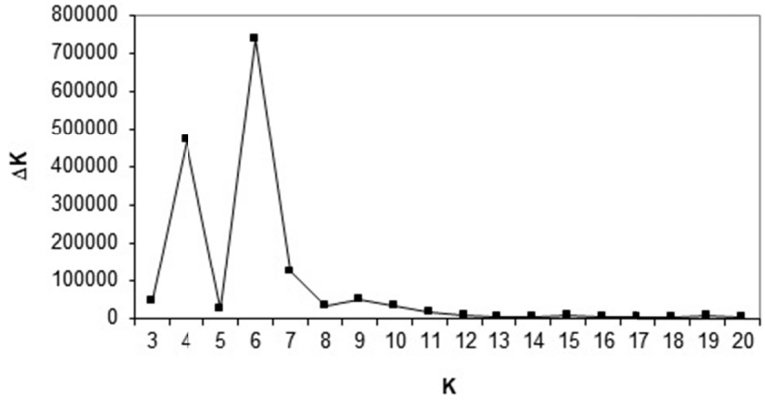


Figure 28. Graphic representation of the ideal number of groups (K = 6) inferred using the methodology of Evano et al. (2005) for the same dataset of Figure 29 (60 rubber trees analyzed with 68 SSRs)

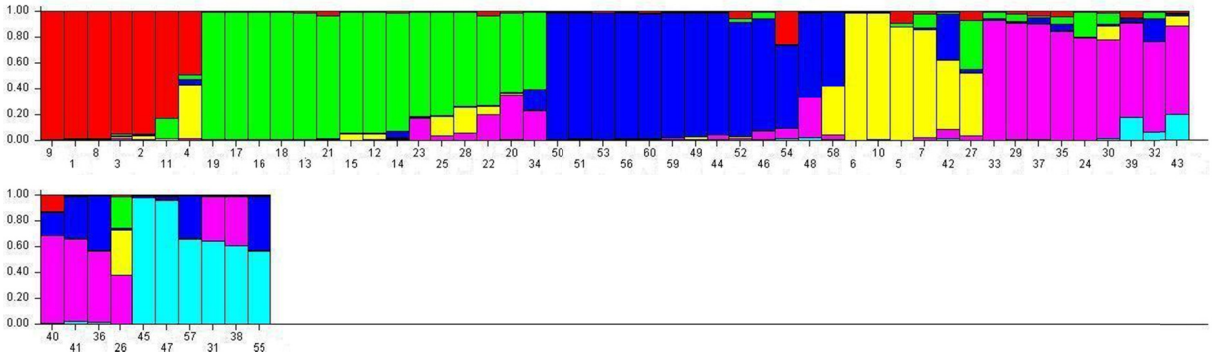


Figure 29. Asian (8, red), African (2, yellow), Amazonian (8, green) and IAC (42, blue) rubber trees analyzed with 68 SSRs

2.2 QTL Mapping

QTLs (Quantitative Trait *Loci*) are chromosome regions responsible for the expression of quantitative traits, *i.e.*, traits with continuous distribution. Some examples are plant height and weight, grain production, oil content, etc. With the development of molecular markers, QTL mapping thrived, since the technique addresses the identification of the genome position of the genes responsible for the trait and estimates the genetic effects, such as additive or dominance effects.

The basic principle of linkage mapping is the existence of linkage disequilibrium (LD) between marker alleles and QTL alleles. What genetic delineations does have LD? From crossing two segregating lines, progenies with different fractions of the genome of each parent are obtained. Several types of segregating populations are in use, most commonly backcrosses, the F_2 generation, pure recombinant inbred lines (RILs) and double-haploid lines obtained from gametes of F_1 plants.

With data from segregating populations generated by specific designs and with individuals analyzed using molecular markers, the next step is the construction of the genetic linkage map. This step encompasses: (a) a segregation (Chi-square) test to check the segregation pattern, choosing only the markers with a Mendelian segregation pattern; (b) linkage test, to verify the linkage disequilibrium between marker pairs with Mendelian segregation, forming a set for each chromosome and (c) ordering of markers based on the distances between them. To form the linkage groups, a statistical test must be applied to check whether the value of the recombination frequency differs from 0.5 by the likelihood ratio (LR), or equivalent tests such as the index known as LOD-score (Logarithm of Odds, Lynch, & Walsh, 1998).

There are many mapping functions that relate distance measurements between the two *loci* on the map to recombination frequencies, among which the most used for plant genetic mapping are the functions of Haldane (1919) and Kosambi (1944). While the first assumes a lack of interference, the second assumes a moderate degree of interference. Interference is caused by the possible occurrence of crossing-over; as the distance

between the *loci* increases, the possibility of this event increases, reducing the accuracy of recombination frequency to estimate the distance (Lynch & Walsh, 1998; Griffiths, Miller, Suzuki, Lewontin, & Gelbart, 2000).

Once segregation is evaluated and linkage groups established by recombination frequencies between marker pairs, the next phase involves measuring distances and ranking within each linkage group. With the increase in the number of markers used in the construction of maps and the increase in the number of individuals analyzed, algorithms and computational programs were developed for the construction of genetic maps, such as Mapmaker/EXP (Lander et al., 1987; Lincoln, Daly, & Lander, 1992), JoinMap (Stam, 1993) and Onemap (Margarido, de Souza, & Garcia, 2007).

Figure 30 shows that each point will be at a certain distance from each of the marker. The greater the distance, the less likely the marker is detecting the effects of that point. The probability of recombination increases with distance. We calculated the effects detected for each marker for each point and multiplied the value by a function that is inversely proportional to the distance between the point and the marker.

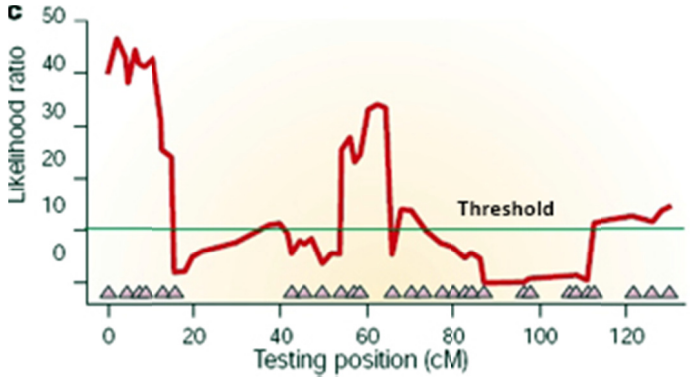


Figure 30. QTL detected and its significance threshold

QTL detection analyses have been advanced towards a higher accuracy of statistical tests and consequently, a greater reliability of the results. Initially, QTLs were identified by a direct relationship between the genotype of the markers of individuals in the mapping population, with their phenotypic characteristics. This method is called Individual Marker Analysis (IMA) and may be based on the maximum likelihood method, or even on linear regression analyses. In this case, scores are attributed to the markers genotype and a simple regression analysis (Equation 2) is performed between these scores (independent variable) in relation to the phenotypic value of the trait (dependent variable). The significant effect of linear regression could identify the existence of an association between the marker and the trait (Schuster & Cruz, 2004).

$$y_j = \mu + \beta x_j + \varepsilon_j \quad (2)$$

where, $j = 1, 2, \ldots n$;

y_j = phenotypic value of the j genotype;

μ = intercept;

$x_j = \begin{cases} 1 \text{ if the genotype is Mi/Mi} \\ 0 \text{ if the genotype is Mi/mi} \end{cases}$;

β = linear regression coefficient (genetic effect);

ε_j = residue ~ $N(0, \sigma^2)$.

By IMA, additive and dominance effects can be estimated; however, it is important to consider a certain redundancy between the "x" variables. For this reason, backward, forward, and stepwise variable selection methods are used, so that a model is sought until that includes all xj variables that are significant by the t test (Bearzoti, 2000). In addition, IMA is conservative, *i.e.*, the number of QTLs that influence the character is not known for sure, aside from not revealing the QTL position, thus underestimating its effect (Ferreira & Grattapaglia, 1998).

Over time, new methodologies have been proposed to overcome the flaws of IMA. The first advance was in the sense of analyzing the existence of QTLs in each interval between two markers, instead of at every single marker, which led to Interval Mapping (Lander & Botstein, 1989). By Interval mapping (IM), the genome is

systematically scanned in search of QTLs, using information from markers flanking the gap, such as recombination fraction and genomic position, to estimate the presence and magnitude of QTLs. In this way, conclusions about the position of the identified QTLs can also be drawn. The significance of this analysis can be tested by regression analysis (Equation 3) or by the maximum likelihood function. In the second case, significance is evaluated by its ratio, by LODscore analysis (Schuster & Cruz, 2004).

$$y_j = \mu + \beta^* x_j^* + \varepsilon_j \quad (3)$$

where, $j = 1, 2, \ldots n$;

y_j = phenotypic value of j genotype;

μ = intercept;

x^*_j $\begin{cases} 1 \text{ if the genotype is QQ} \\ 0 \text{ if the genotype is Qq'} \end{cases}$

β^* = possible effect of the QTL;

ε_j = residue ~ N(0, σ^2).

A disadvantage of this method is that other QTLs outside the range in question are not considered, which has two consequences. The first is that all genetic variation, due to these other QTLs, are residuals, which decreases the precision of the estimates and the power of the test. The second is that if two QTLs are linked in the considered range, inexistent QTLs, known as phantom QTLs, might be identified. To avoid these effects, Jansen (1993) and Zeng (1993) independently proposed methods by which QTLs outside the range in question are considered by the multiple regression method (Bearzoti, 2000).

Then, Zeng (1994) proposed the Compound Interval Mapping (CIM). In this approach, the effects between QTL from other regions do not influence the analysis between two markers. This occurs using the multiple regression method (Equation 4), which reduces the residual variance between the *loci* and thus increases the detection power of each QTL and the accuracy in estimating its effects (Zeng, 1994; Jansen & Stan, 1994). To minimize the effects of other QTLs outside the range under analysis, they are included in it as cofactors, previously determined by stepwise procedure (Zeng, 1994).

With this, a model for each position in the genome can be constructed, testing the significance of the additive and dominance effects of each model by maximum likelihood and LODscore. Since numerous tests are performed for each mapping experiment, the significance in each case can be computed by several methods, of which the permutation test is the most indicated (Churchill & Doerge, 1994).

$$y_j = \mu + \beta^* x_j^* + \sum_k \beta_k x_{jk} + \varepsilon_j \quad (4)$$

where, $j = 1, 2, \ldots n$;

y_j = value c of genotype j;

μ = intercept;

x_j^* = $\begin{cases} 1 \text{ if the genotype of the QTL is QQ} \\ 0 \text{ if the genotype of the QTL is Qq'} \end{cases}$

β^* = effect of the possible QTL;

x_{jk} = cofactors;

ε_j = residue ~ N(0, σ^2).

The estimation of the genetic effects of QTLs is somewhat problematic. Depending on the genetic and experimental design and the models used to estimate them, estimates may be biased. The biases are caused by deficiencies of recombinant gametes, genotype-environment interaction and underestimation of epistasis, mainly if the mapped populations has a small size. The solutions to this impasse would be to increase the number of families to at least 300, increase the density of the map and to analyze only extreme phenotypes (Lee, 1995).

With this in mind, Jiang and Zeng (1995) extended the CIM concept to map multiple QTLs detected in different environments, aiming to study the interaction of pleiotropy between QTLs and genotype-environment interaction. As a result, a method called Multiple Interval Mapping (MIM) was proposed by Kao, Zeng, and Teasdale (1999), which incorporates epistasis into the model and considers multiple intervals simultaneously.

The method consists of the selection and comparison of models, based on four components: evaluation, search, estimation, and prediction. For this purpose, a pre-model is generated for each analysis by multiple regression and cofactor selection, as in the CIM analysis. However, the number of QTLs and possible interactions between

them are now considered analysis parameters (Equation 5), which allows eliminating non-significant results. Thus, the selected model proceeds to the effective QTL analysis with these "new" parameters, defined in real terms by data analysis. This procedure is repeated several times until all insignificant QTLs are excluded from the analyses. Thus, this method is more advantageous due to the greater efficiency and accuracy in the search for QTLs and their interactions.

$$y_j = \mu + \sum_{r=1}^{m} \beta_r \cdot x_{jr}^{*} + \sum_{r \neq s \subset (1,\ldots m)}^{t} \beta_{rs} (x_{jr}^{*} \cdot x_{js}^{*}) + \varepsilon_j \quad (5)$$

where, $j = 1, 2, \ldots n$;

y_j = phenotypic value of the j genotype;

μ = intercept;

β_r = marginal effect of the possible QTL_r;

x_{jr}* = indicator variable of the genotype of the possible QTL_r;

β_{rs} = epistatic effect between r and s QTLs;

ε_j = residue ~ $N(0, \sigma^2)$.

Currently, in view f the advantages, the MIM method has been most widely indicated for the search of new QTLs, mainly due to its statistical precision with more efficient results, including the use of these QTLs in marker-assisted studies. However, the most commonly used method is still CIM, which has also met the interests of researchers adequately in several studies (Maxwell et al., 2007; Sabadin, Souza Júnior, Souza, & Garcia, 2008; Figures 31 and 32).

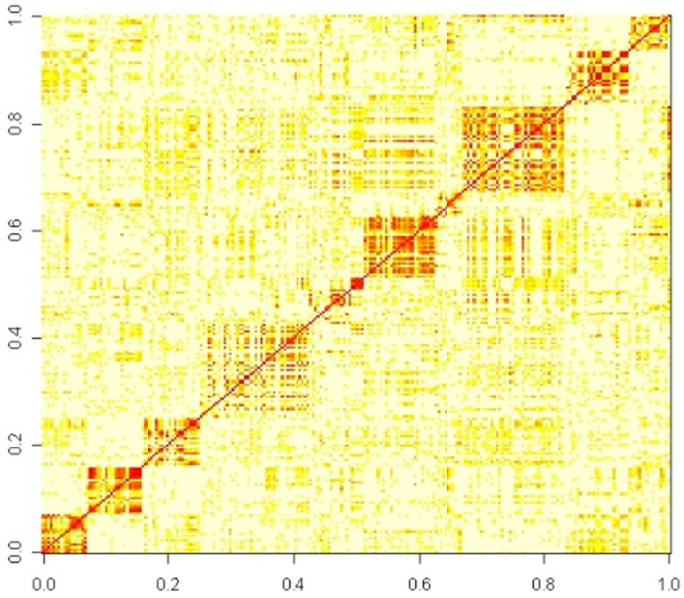


Figure 31. Recombination fraction plot between *loci* of the 11 linkage groups of common bean identified by Two-Point analysis. Oblessuc et al. (2012)

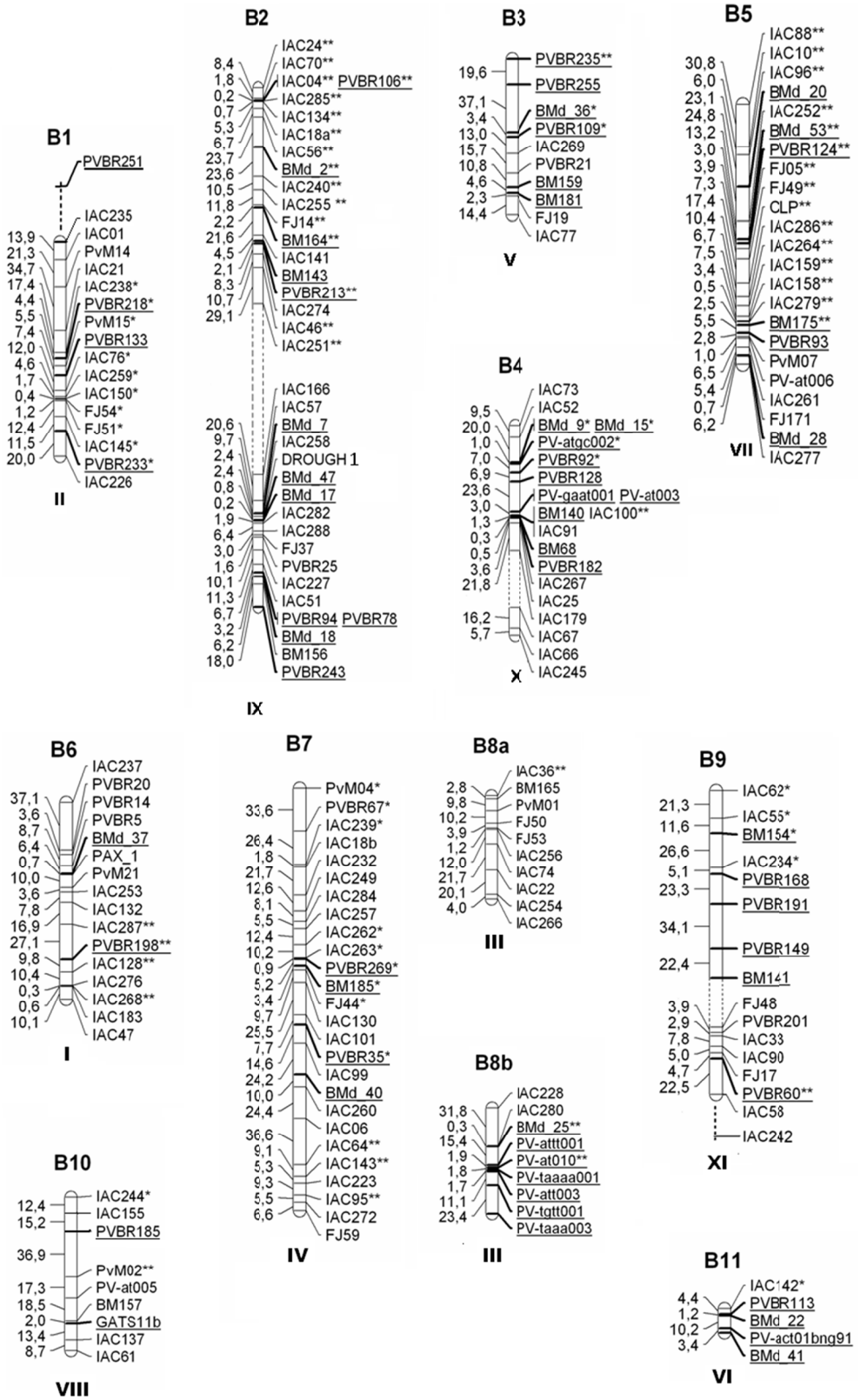


Figure 32. Genetic map for common bean, derived from linkage analyses between genotyped microsatellites in the population 'IAC-UNA' × 'CAL 143' (UC). The *loci* previously mapped in other populations and anchored in their respective linkage groups for two-point analysis (LOD ≥ 3.0 and r ≥ 0.40) are underlined. (*) *loci* with segregation deviation for the Andean parent allele ('CAL 143'). (**) *loci* with segregation deviation for the Mesoamerican parent ('IAC-UNA'). 5% significance level. Oblessuc et al. (2012)

2.3 Associative Mapping and Linkage Disequilibrium

Genetic mapping can be performed mainly in two ways (1) using experimental populations (also called a "biparental" population or population mapping), which is known as QTL mapping or linkage mapping, and (2) using several genotypes of natural populations or germplasm collections, which is called LD mapping or associative mapping. Traditionally, QTL mapping approaches have been based on the analysis of populations derived from biparental crosses segregating for a trait of interest. There are two strategies that can be used in associative mapping: a) association mapping by already identified candidate genes (Candidate Gene Approach); b) genome-wide association studies (GWAS), using molecular markers to cover the entire genome to identify regions that are associated with a phenotype of interest.

Associative mapping is an alternative method that can be used to associate a molecular marker to a trait of interest in natural populations or in a collection of cultivars of a breeding program (Oraguzie, Rikkerink, Gardiner, & Silva, 2007). The underlying principle of this approach is that LD occurs, since this is necessary for the association between marker and phenotype, and because the extent of these regions will determine the resolution of the map. Linkage disequilibrium is also defined as the non-random association of different *loci* (Flint-Garcia, Thornsberry, & Buckler, 2003). This imbalance is attributed to the physical connection between the *loci*, which alters the expected frequency of connections, generating non-random recombinations within the chromosome, which makes it possible to detect the connection between them (Resende, 2008).

A relative measure of LD, called *D*' (Equation 6), was proposed by Lewontin (1964), by which the LD of different *locus* pairs can be compared by taking the maximum theoretical values into consideration, according to the formula:

$$D' = \frac{|D|}{D_{max}} \tag{6}$$

where: *D*' is the relative measure of LD; *D* the basic concept calculated based on haplotype frequencies; and D_{max} the maximum theoretical value of LD between a possible pair of *loci*. In contrast to the basic concept of LD, which can vary between -0.25 and 0.25, *D*' can vary between 0 and 1.

Another relative measure of LD is r^2. This measure was proposed by Hill & Robertson (1968) and consists of the square of the correlation coefficient. It measures the degree of association between *loci* (covariance), according to the variation of their alleles. As the measure of *D*', r^2 depends on the haplotype frequencies and is calculated as follows (Equation 7):

$$r_{XY}^2 = \frac{Cov(X,Y)}{V(X) \cdot V(Y)} = \frac{D_{XY}^2}{f_X \cdot f_x \cdot f_Y \cdot f_y} = \frac{D_{XY}^2}{f_X(1 - f_X) \cdot f_Y(1 - f_Y)} \tag{7}$$

where: r_{XY}^2 (coefficient of determination, or square of the correlation coefficient) is the relative measure of LD between any two biallelic *X* and *Y loci*, *Cov*(*X*,*Y*) is the covariance between these *loci*, *V*(*X*) and *V*(*Y*) are their respective variances, D_{XY}^2 is the basic concept of LD between *X* and *Y*, and $f_{X,x,Y,y}$ are the allele frequencies referring to *X* and *Y loci* ($X: f_X + f_x = 1$), $V(X) = f_X \cdot f_x$; $Y: f_Y + f_y = 1$; $V(Y) = f_Y \cdot f_y$.

Similarly to *D*', r^2 can vary between 0 and 1. The measures *D*' and r^2 are the most commonly used to calculate the LD between pairs of biallelic *loci*. Although both are not suitable for measuring LD in small samples, with or without low allele frequencies, each has its advantages. While r^2 capitalizes on historically occurring mutation and recombination events, *D*' capitalizes on recombination events only, and is the most adequate measure to detect them. However, *D*' is strongly affected by small samples, so that comparisons between *loci* with low allele frequencies with this measure are inappropriate due to the high bias of the LD estimates (Flint-Garcia, Thornsberry, & Buckler, 2003). Therefore, in the case of association studies, where the sample size is limited by time and cost issues involved in the genotyping and phenotyping processes, r^2 should be used to verify the extent of LD.

Several aspects influence the LD observed in a population or species: mutation and recombination rate, mating system, genetic drift, population structure, kinship degree, selection, epistasis, and chromosomal rearrangements (Abdurakhmonov & Abdukarimov, 2008). These factors can generate false associations between markers and phenotypes in associative mapping, so they must be considered in the analyses.

Currently, on account of the advancement of NGS, the GWAS has become the most widely used approach. Phenotypic and genotypic data of many individuals are assembled in a diversity panel. Genotyping data usually consists of genome-wide single-nucleotide polymorphisms (SNPs) identified by resequencing, genotyping by sequencing (GBS) or genotyping based on an array containing SNPs (BeadChips, Illumina). As GWAS uses

unrelated individuals with greater genetic variability, the number of markers must be high, which is possible due to the development of high-resolution genome analysis technology with NGS.

Genome-wide association studies identify variations in the genome and associates them with the phenotype of interest by hypothesis testing, to construct genetic maps. The primary objective of GWAS is to identify marker-trait associations for one trait at a time and, secondly, to study its genetic architecture. The latter involves identifying all QTL/genes. The extent and level of this information has also improved with the continued increase in size of the association panels and the number of molecular markers used for GWAS. The identification of many false positives that appear after GWAS analysis and false negatives detected by Bonferroni or false discovery rate (FDR) corrections has been a problem in GWAS. The level of significance must be carefully considered, since, as thousands of markers are tested, setting the level at 5%, as generally done in the various tests, could boost the false-positive rate. The population structure or existence of epistasis between *loci* can also cause false positives (Cortes, Zhang, & Yu, 2021).

2.4 Marker-assisted selection in Backcrosses

With the advent of molecular markers and the first genetic molecular maps, marker-assisted selection (MAS, Figure 33) became an attractive idea for breeders. In backcrosses, molecular marker technology can be applied in several stages. The best parents for a cross can be identified by DNA fingerprinting of germplasm collections; the divergence between the parents can also be determined, to estimate the effort that will be required to re-establish the elite parent genotype. In each backcross cycle, the markers can be used to identify lines carrying the target allele (de Almeida et al., 2021; Paulino et al., 2022), containing the least proportion of the donor genotype, and to find lines with the least number of segments around the target allele (linkage drag). For recessive traits, molecular markers may prevent the need for an additional generation of selfing to identify the target allele (Langridge et al., 2001).

Openshaw, Jarboe, and Beavis (1994) reported that the number of backcross generations in maize lines could be reduced from seven to three if the sample size were smaller ($N < 100$) and less markers were used ($N < 80$). Hospital, Chevalet & Mulsant (1992) concluded that, with the use of MAS, the number of generations could be reduced from seven to two. For most crops, more than 90% of the recurrent parent genotype can be recovered within two generations, if an adequate number of markers and an adequate number of progenies are used for genome selection (Tanksley, Young, Paterson, & Bonier-bale, 1989). According to Benchimol, de Souza Jr., and de Souza (2005), the means of recovery of recurrent genotypes in three backcross generations were compatible with those expected in BC_4 or BC_5, indicating genetic gain due to marker-assisted backcrossing. The formula below estimates the percentage of the recurrent genome still in the donor's genome (Equation 8):

$$GR\%* = [B + (0.5H)/(B + H + A)] \times 100 \quad (8)$$

where: A is the genome of the donor parent, B the genome of the recurrent parent and H the F_1 hybrid.

Marker-assisted backcross efficiency is based on various factors such as population size, distance of markers from the target locus and number of background markers used (Hasan et al., 2015). According to Kim et al. (2021), line selection based on KASP (Kompetitive allele-specific PCR) markers was successful in BC_1F_1 and BC_2F_1, with a recovery of 97-99.1% of the recurrent parent genome.

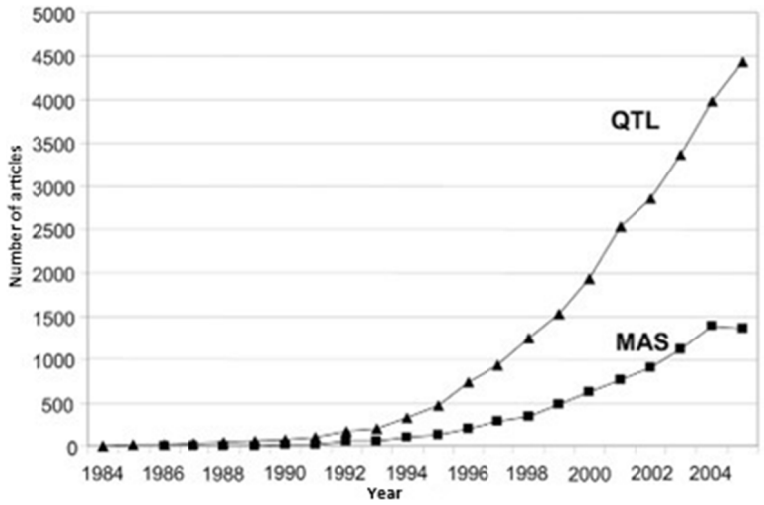


Figure 33. Number of articles with the terms QTL (Quantitative trait *loci*) and marker-assisted selection (MAS) per year (1984-2005). Xu and Crouch (2008)

2.5 Genome-Wide Selection

Genome-wide selection (GWS), proposed by Meuwissen, Hayes, and Goddard (2001), is an analysis based on the evaluation of many markers widely distributed throughout the genome, only feasible with the development and cost reduction of SNP-type markers. Once these markers are available, their effects are estimated based on phenotypic data from an estimated population. After estimating the effects, the prediction models of genomic genetic values (GGV) are obtained. Model accuracy is achieved from validation populations and then applied to selection populations (Crossa et al., 2011).

In breeding, once the genomic genetic values are predicted, they can be used to evaluate the genetic values of individuals in early selection stages, with the prediction of future phenotypes of individuals genotyped in preliminary stages of the breeding program. This strategy speeds up selection, considering that in certain selection cycles, genotypes of interest can be selected without the need for phenotyping (Resende, Lopes, Silva, & Pires, 2008).

In GWS, a LD between the QTL and the marker is required, and only markers in LD can explain the genetic variation. As a wide range of markers is used, the probability of finding a QTL in LD with the marker is high. This technique is superior to MAS because it is highly accurate, can be applied to a larger number of families within the population, and does not require prior knowledge of the QTL positions on a map (Miqueloni, Simeão, & de Assis, 2019). Marker-assisted selection only captures major-effect QTLs, but cannot explain all genetic variation of the individuals, whereas GWS detects both major and minor-effect QTLs

Three types of population sets are required for the implementation of GWS:

- Estimation population, where many SNPs are tested in a moderate number of individuals already phenotypically characterized and a prediction equation of genomic values, which uses the markers as input data, is generated.

- Validation population, smaller than the first, where individuals are phenotypic and genetically characterized for markers and prediction equations are tested to assess accuracy in this independent sample.

- Selection population, where individuals are genetically characterized and prediction equations estimated in the first set, used to calculate the prediction methods of genomic values, but considering the accuracy of the validation set (Figure 34).

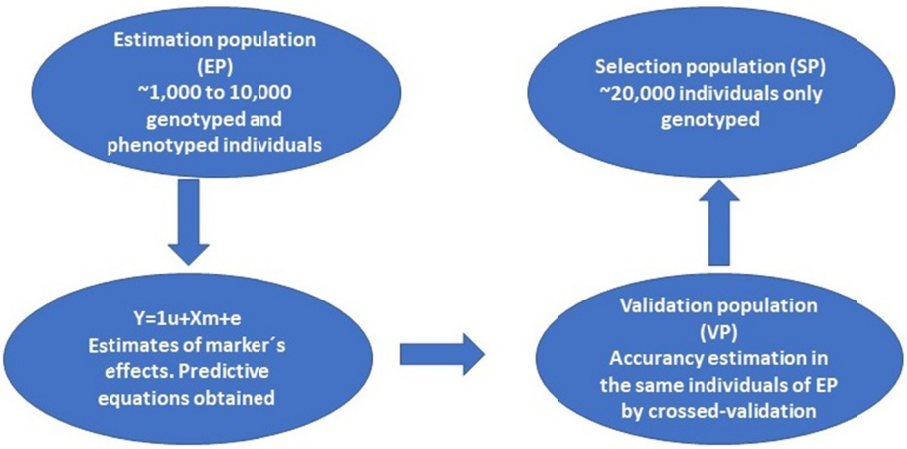


Figure 34. Diagram of genome-wide selection in a breeding program
(Resende et al., 2010)

In cross-validation, the population is divided into k groups of the same size and in each of the groups, k individuals are removed to form the validation population. In each subgroup, the predicted genetic value and phenotypic value of the evaluated individuals are correlated (Pereira, 2021). The GWS was first used in plants, mainly to select quantitative inheritance traits, with low heritability. Genome-wide selection prioritizes the prediction of genotype performance and not the genetic architecture of quantitative traits.

Due to the large number of effects to be estimated simultaneously from a small number of observations, the sample size, collinearity of the LD between the markers and the interpretation of the results, adequate statistical models and methodologies are needed to predict the marker effects (Pereira, 2021). The GBLUP (Genomic Best Linear Unbiased Prediction) is one of the most commonly used models for GWS, with a single normal distribution of the markers (Li, Wang, & Bao, 2015).

3. Conclusion

With the advances in NGS a flourishing use of new molecular marker systems is to be expected. However, old techniques are still in use and the choice will be based on the research objectives and funding of each project. Single Nucleotide Polymorphisms have been extensively explored in plant breeding. However, they are not preferable to SSRs with regard to detecting fundamental genetic problems such as gene flow, apparent outcrossing rates, pleiotropy and/or epistasis. Genome-wide selection is being considered superior to MAS if quantitative traits are targeted; however, MAS is still being widely applied, mostly in plant breeding companies. QTL-mapping and genome-wide association studies (GWAS) are being widely used to decode the genetic basis of polygenic traits.

References

Abdurakhmonov, I. Y., & Abdukarimov, A. (2008). Application of Association Mapping to Understanding the Genetic Diversity of Plant Germplasm Resources. *International Journal of Plant Genomics*, 574927. https://doi.org/10.1155/2008/574927

Batley, J., Mogg, R., Edwards, D., Osullivan, H., & Edwards, K. J. (2003). A high-throughput SNuPE assay for genotyping SNPs in flanking regions of *Zea mays* sequence tagged simple sequence repeats. *Molecular Breeding, 11*, 111-120. https://doi.org/10.1023/A:1022446021230

Bearzoti, E. (2000). Mapeamento de QTL. In J. B. Pinheiro & I. F. Carneiro (Eds.), *Análise de QTL no melhoramento de plantas* (pp. 63-223). Goiânia: FUNAPE.

Benchimol, L. L., de Souza Jr., C. L., & de Souza, A. P. (2005). Microsatellite-assisted backcross selection in maize. *Genetics and Molecular Biology, 28*(4), 789-797. https://doi.org/10.1590/S1415-47572005000500022

Botstein, D., White, R. L., Skolnick, M., & Davis, R. W. (1980). Construction of genetic linkage map using restriction fragment length polymorphisms. *The American Journal of Human Genetics, 32*(3), 314-31.

Briñez, B., Blair, M., Kilian, A., Carbonell, S. A. M., Chiorato, A. F., & Rubiano, L. B. (2012). A whole genome DArT assay to assess germplasm collection diversity in common beans. *Molecular Breeding, 30*, 181-193. https://doi.org/10.1007/s11032-011-9609-3

Chen, Y., Peng, Z., Wu, C., Ma, Z., Ding, G., Cao, G., … Lin, S. (2017). Genetic diversity and variation of Chinese fir from Fujian province and Taiwan, China, based on ISSR markers. *Plos One, 12*, 01-14. https://doi.org/10.1371/journal.pone.0175571

Churchill, G. A., & Doerge, R. W. (1994). Empirical threshold values for quantitative trait mapping. *Genetics, 138*(3), 963-71. https://doi.org/10.1093/genetics/138.3.963

Collard, B. C., Jahufer, M. Z., Brouwer, J. B., & Pang, E. C. K. (2005). An introduction to markers, quantitative trait *loci* (QTL) mapping and marker-assisted selection for crop improvement: the basic concepts. *Euphytica, 142*(1-2), 169-196. https://doi.org/10.1007/s10681-005-1681-5

Cortes, L. T., Zhang, Z., & Yu, J. (2021). Status and prospects of genome-wide association studies in plants. *The Plant Genome, 14*(1), e20077. https://doi.org/10.1002/tpg2.20077

Costa, R., Pereira, G., Garrido, I., Tavares-de-Sousa, M. M., & Espinosa, F. (2016). Comparison of RAPD, ISSR, and AFLP molecular markers to reveal and classify orchardgrass (*Dactylis glomerata* L.) germplasm variations. *PLoS One, 11*, e0152972. https://doi.org/10.1371/journal.pone.0152972.

Crossa, J., Pérez, P., Campos, G. de L., Mahuku, G., Dreisigacker, S., & Magorokosho, C. (2011). Genomic Selection and Prediction in Plant Breeding. *Journal of Crop Improvement, 25*(3), 239-261. https://doi.org/10.1080/15427528.2011.558767

De Almeida, C. P., Paulino, J. F. D. C., Santos, I. L., Bajay, M. M., Gonçalves, G. D. M. C., Carvalho, C. R. L., ... Benchimol-Reis, L. L. (2021). Marker-assisted backcrossing for disease resistance and agronomic traits in Carioca beans. *Crop Science, 61*(4), 2510-2521. https://doi.org/10.1002/csc2.20528

Drenkard, E., Richter, B. G., Rozen, S., Stutius, L. M., Angell, N. A., Mindrinos, M., … Ausubel, F. M. (2000). A simple procedure for the analysis of single-nucleotide polymorphisms facilitates map-based cloning in Arabidopsis. *Plant Physiology, 124*, 1483-1492. https://doi.org/10.1104/pp.124.4.1483

Elshire, R. J., Glaubitz, J. C., Sun, Q., Poland, J. A., Kawamoto, K., Buckler, E. S., & Mitchell, S. E. (2011). A Robust, Simple Genotyping-by-Sequencing (GBS) Approach for High Diversity Species. *PLoS ONE, 6*(5), e19379. https://doi.org/10.1371/journal.pone.0019379

Evano, G., Regnaut, S., & Goudet, J. (2005). Detecting the number of clusters of individuals using the software structure: A simulation study. *Molecular Ecology, 14*, 2611-2620. https://doi.org/10.1111/j.1365-294X.2005.02553.x

Ferreira, M. E., & Grattapaglia, D. (1995). *Introdução ao Uso de Marcadores RAPD, RFLP em Análise Genética* (p. 220). Brasília: EMPRAPA-CENARGEN.

Flint-Garcia, S. A., Thornsberry, J. M., & Buckler, I. V. (2003). Structure of linkage disequilibrium in plants. *Annual Review in Plant Biology, 54*, 357-374. https://doi.org/10.1146/annurev.arplant.54.031902.134907

Goodman, M. M., & Stuber, C. W. (1983). Races of maize VI. Isozyme variation among races of maize in Bolívia. *Maydica, 28*, 169-187.

Griffiths, A. J. F., Miller, J. H., Suzuki, D. T., Lewontin, R. C., & Gelbart, W. M. (2000). *An introduction to genetic analysis* (pp. 1-103). New York: W. H. Freeman.

Haldane, J. B. S. (1919). The combination of linkage values, and the calculation of distance between the loci of linked factors. *Journal of Genetics, 8*, 299-309. https://doi.org/10.1007/BF02983270

Hamada, H., Petrino, M., & Kakunaga, T. (1982). A novel repeated element with Z-DNA-forming potential is widely found in evolutionarily diverse eukaryotic genomes. *Proceedings of the National Academy of Sciences, 79*, 646-649. https://doi.org/10.1073/pnas.79.21.6465

Hasan, M. M., Rafii, M. Y., Ismail, M. R., Mahmood, M., Rahim, H. A., Alam, M. A., … Latif, M. A. (2015). Marker-assisted backcrossing: A useful method for rice improvement. *Biotechnology and Biotechnological Equipment, 29*(2), 237-254. https://doi.org/10.1080/13102818.2014.995920

Hill, W. G., & Robertson, A. (1968). Linkage disequilibrium in finite populations. *Theoretical and Applied Genetics, 38*, 226-231. https://doi.org/10.1007/BF01245622

Hospital, F., Chevalet, C., & Mulsant, P. (1992). Using markers in gene introgression breeding programs. *Genetics, 132*, 1199-1210. https://doi.org/10.1093/genetics/132.4.1199

Hu, J., & Vick, B. A. (2003). Target Region Amplification Polymorphism: A Novel Marker Technique for Plant Genotyping. *Plant Molecular Biology Reporter, 21*, 289-294. https://doi.org/10.1007/BF02772804

Hwang, M. T., Landon, P. B., Lee, J., Choi, D., Mo, A. H., Glinsky, G., & Lal, R. (2016). Highly specific SNP detection using 2D graphene electronics and DNA strand displacement. *Proceedings of the National Academy of Sciences, 113*(26), 7088-7093. https://doi.org/10.1073/pnas.1603753113

Jaccoud, D., Peng, K., Feinstein, D., & Kilian, A. (2001). Diversity arrays: A solid-state technology for sequence information independent genotyping. *Nucleic Acids Research, 29*, e25. https://doi.org/10.1093/nar/29.4.e25

Jansen, R. C. (1993). Interval Mapping of Multiple Quantitative Trait *Loci*. *Genetics, 135*, 205-211. https://doi.org/10.1093/genetics/135.1.205

Jansen, R. C., & Stam, P. (1994). High Resolution of Quantitative Traits into Multiple *Loci* via Interval Mapping. *Genetics, 136*, 1447-1455. https://doi.org/10.1093/genetics/136.4.1447

Jiang, C., & Zeng, Z. B. (1995). Multiple Trait Analysis of Genetic Mapping for Quantitative Trait *Loci*. *Genetics, 140*, 1111-1127. https://doi.org/10.1093/genetics/140.3.1111

Kalia, R. K., Rai, M. K., Kalia, S., Singh, R., & Dhawan, A. K. (2011). Microsatellite markers: An overview of the recent progress in plants. *Euphytica, 177*, 309-334. https://doi.org/10.1007/s10681-010-0286-9

Kao, C. H., Zeng, Z. B., & Teasdale, R. D. (1999). Multiple interval mapping for quantitative trait *loci*. *Genetics, 152*, 1203-1216. https://doi.org/10.1093/genetics/152.3.1203

Karaköy, T., Baloch, F. S., Toklu, F., & Özkan, H. (2014). Variation for selected morphological and quality-related traits among 178 faba bean landraces collected from Turkey. *Plant Genet Resources, 12*(01), 5-13. https://doi.org/10.1017/S1479262113000208

Kim, M. S., Yang, J. Y., Yu, J. K., Lee, Y., Park, Y. J., Kang, K. K., & Cho, Y. G. (2021). Breeding of High Cooking and Eating Quality in Rice by Marker-Assisted Backcrossing (MABc) Using KASP Markers. *Plants (Basel), 19, 10*(4), 804. https://doi.org/10.3390/plants10040804

Kosambi, D. D. (1944). The estimation of map distance from recombination values. *Annuaire of Eugenetics, 12*, 172-175. https://doi.org/10.1111/j.1469-1809.1943.tb02321.x

Lander, E. S., & Botstein, D. (1989). Mapping Mendelian Factors Underlying Quantitative Traits Using RFLP Linkage Maps. *Genetics, 121*, 185-199. https://doi.org/10.1093/genetics/121.1.185

Lander, E. S., Green, P., Abrahamson, J., Barlow, A., Daly, M. J., Sen, L., & Newburg, L. (1987). Mapmaker: an interactive computer package for constructing primary genetic linkage maps of experimental and natural populations. *Genomics, 1*, 174-181. https://doi.org/10.1016/0888-7543(87)90010-3

Langridge, P., Lagudah, E. S., Holton, T. A., Appels, R., Sharp, P. J., & Chalmers, K. J. (2001). Trends in genetic and genome analyses in wheat: a review. *Australian Journal of Agricultural Research, 52*, 1043-1077. https://doi.org/10.1071/AR01082

Lee, M. (1995). DNA markers and plant breeding program. *Advances in Agronomy, 55*, 265-344. https://doi.org/10.1016/S0065-2113(08)60542-8

Lewontin, R. C. (1964). The interaction of selection and linkage. I. General considerations; heterotic models. *Genetics, 49*, 49-67. https://doi.org/10.1093/genetics/49.1.49

Li, H., Wang, J., & Bao, Z. (2015). A novel genomic selection method combining GBLUP and LASSO. *Genetica, 143*(3), 299-304. https://doi.org/10.1007/s10709-015-9826-5

Lincoln, S. E., Daly, M. J., & Lander, E. S. (1992). *Constructing genetic maps with Mapmaker Exp 3.0* (p. 230, 3rd ed.). Cambridge: Whitehead Institute for Biometrical Research.

Lynch, M., & Walsh, B. (1998). *Genetics and analysis of quantitative traits* (p. 980). Sinauer Associates, Sunderland, MA.

Margarido, G., de Souza, A., & Garcia, A. (2007). OneMap: Software for genetic mapping in outcrossing species. *Hereditas, 144*(3), 78-79. https://doi.org/10.1111/j.2007.0018-0661.02000.x

Maxwell, J. J., Brick, M. A., Byrne, P. F., Schwartz, H. F., Shan, X., Ogg, J. B., & Hensen, R. A. (2007). Quantitative Trait Loci Linked to White Mold Resistance in Common Bean. *Crop Breeding & Genetics, 47*, 2285-2294. https://doi.org/10.2135/cropsci2007.01.0022

Meuwissen, T. H., Hayes, B. J., & Goddard, M. E. (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics, 157*(4), 1819-1829. https://doi.org/10.1093/genetics/157.4.1819

Miller, M. (1997). A Windows program for the analysis of allozyme and molecular population genetic data. *Tools for Population Genetic Analysis (TFPGA).*

Miqueloni, D. P., Simeão, R. M., & de Assis, G. M. L. (2019). Seleção genômica ampla (GWS) e associação genômica ampla (GWAS) no melhoramento de forrageiras: Abordagem conceitual, genética quantitativa e aplicações. *Enciclopédia Biosfera, Centro Científico Conhecer, Goiânia, 16*(30), 556-582. https://doi.org/10.18677/EnciBio_2019B52

Mullis, K. B., & Faloona, F. A. (1987). Specific synthesis of DNA in vitro via a polymerase-catalyzed chain reaction. *Methods in Enzymology, 155*, 335-350. https://doi.org/10.1016/0076-6879(87)55023-6

Nelson, M. R., Marnellos, G., Kammerer, S., Hoyal, C. R., Shi, M. M., Cantor, C. R., & Braun, A. (2004). Large-scale validation of single-nucleotide polymorphisms in gene regions. *Genome Research, 14*(8), 1664-1668. https://doi.org/10.1101/gr.2421604

Oblessuc, P. R., Baroni, R., Garcia, A. A. F., Chiorato, A. F., Carbonell, S. A. M., Camargo, L. E. A., & Benchimol, L. L. (2012) Mapping of angular leaf spot resistance QTL in common bean (*Phaseolus vulgaris* L.) under different environments. *BMC Genetics, 13*, 50. https://doi.org/10.1186/1471-2156-13-50

Oliveira, E. J., Pádua, J. G., Zucchi, M. I., Vencovsky, R., & Vieira, M. L. C. (2006). Origin, evolution and genome distribution of microsatellites. *Genetics and Molecular Biology, 29*(2), 294-307. https://doi.org/10.1590/S1415-47572006000200018

Openshaw, S. J., Jarboe, S. G., & Beavis, W. D. (1994). *Marker-assisted selection in backcross breeding* (pp. 41-43). Proceedings of the Symposium "Analysis of Molecular Marker Data", Joint Plant Breeding Symposia Series, American Society for Horticultural Science/Crop Science of America, Corvallis, Oregon.

Oraguzie, N. C., Rikkerink, E. H. A., Gardiner, S. E., & Silva, H. N. (2007). *Association mapping in plants* (p. 277). New York: Springer. https://doi.org/10.1007/978-0-387-36011-9

Paran, I., & Michelmore, R. W. (1993). Development of Reliable PCR-Based Markers Linked to Downy Mildew Resistance Genes in Lettuce. *Theoretical and Applied Genetics, 85*, 985-993. https://doi.org/10.1007/BF00215038

Paulino, J. F. C., Almeida, C., Santos, I. L., Gonçalves, J. G. R., Carbonell, S. A. M., Chiorato, A. F., & Benchimol-Reis, L. L. (2022). Combining disease resistance and postharvest quality traits by early marker-assisted backcrossing in carioca beans. *Scientia Agricola, 79*(2). https://doi.org/10.1590/1678-992X-2020-0233

Penteado, M. I. D. O., García, P. & de la Veja, M. P. (1997). Isozyme markers and genetic variability in three species of *Centrosema* (Leguminosae). *Brazilian Journal of Genetics*, 20 (3). https://doi.org/10.1590/S0100-84551997000300015

Pereira, G. C. (2021). *Seleção genômica ampla (GWS) sob assimetria para resistência à podridão da espiga em milho* (p. 42, Unpublished Masters thesis, Federal University of Lavras, Minas Gerais, Brazil).

Pritchard, J. K., & Wen, W. (2004). *Documentation for structure software: Version 2.1*. Retrieved from http://pritch.bsd.uchicago.edu

Pritchard, J. K., Stephens, M., & Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics, 155*, 945-959. https://doi.org/10.1093/genetics/155.2.945

Rafalski, A. (2002). Applications of single-nucleotide polymorphisms in crop genetics. *Current Opinion in Plant Biology, 5*(2), 94-100. https://doi.org/10.1016/s1369-5266(02)00240-6

Resende, M. D. V. de. (2008). *Genômica quantitativa e seleção no melhoramento de plantas perenes e animais* (p. 330). Colombo: Embrapa Florestas. Retrieved from http://livimagens.sct.embrapa.br/amostras/00051790.pdf

Resende, M. D. V., Lopes, O. S., Silva, R. L., & Pires, I. E. (2008). Seleção genômica ampla (GWS) e maximização da eficiência do melhoramento genético. *Pesquisa Florestal Brasileira, 1*(56), 63-77. Retrieved from https://pfb.cnpf.embrapa.br/pfb/index.php/pfb/article/view/63

Resende, M. D. V., Resende Júnior, M. F. R., Aguiar, A. M., Abad, J. I. M., Missiaggia, A. A., Sansaloni, C., ... Grattapaglia, D. (2010). *Computação da Seleção Genômica Ampla (GWS)* (p. 79). Colombo, PR: Embrapa Florestas. Retrieved from https://www.infoteca.cnptia.embrapa.br/handle/doc/883425

Rohlf, F. J. (1997). *NTSYS-pc Version 2.02i Numerical Taxonomy and Multivariate Analysis System*. Applied Biostatistics Inc., Exeter Software, Setauket, New York.

Sabadin, P. K., Souza Júnior, C. L., Souza, A. P., & Garcia, A. A. F. (2008). QTL mapping for yield components in a tropical maize population using microsatellite markers. *Hereditas, 145*, 194-203. https://doi.org/10.1111/j.0018-0661.2008.02065.x

Schlötterer, C., & Tautz, D. (1992). Slippage synthesis of sample sequence DNA. *Nucleic Acids Research, 20*, 211-215. https://doi.org/10.1093/nar/20.2.211

Schuster, I., & Cruz, C. D. (2004). *Estatística genômica—Aplicada a Populações Derivadas de Cruzamentos Controlados* (p. 568). Universidade Federal de Viçosa, Brasil. https://doi.org/10.12702/1984-7033.v05n02a19

Sibov, S. T., Souza Jr., C. L., Garcia, A. A. F., Garcia, A. F., Silva, A. R., Mangolin, C. A., ... Souza, A. P. (2003). Molecular mapping in tropical maize (*Zea mays* L.) using microsatellites markers. Part 1: Map construction and localization of *loci* showing distorted segregation. *Hereditas, 139*, 96-106. https://doi.org/10.1111/j.1601-5223.2003.01666.x

Stam, P. (1993). Construction of integrated genetic maps by means of a new computer package: JoinMap. *The Plant Journal, 3*, 739-744. https://doi.org/10.1111/j.1365-313X.1993.00739.x

Tanksley, S. D., Young, N. D., Paterson, A. H., & Bonier-bale, M. W. (1989). RFLP mapping in plant breeding: New tools for an old science. *Biotechnology, 7*, 257-263. https://doi.org/10.1038/nbt0389-257

Tautz, D., & Renz, M. (1984). Simple sequences are ubiquitous repetitive components of eukaryotic genomes. *Nucleic Acids Research, 12*, 4127-4138. https://doi.org/10.1093/nar/12.10.4127

Tivang, J. G., Nienhuis, J., & Smith, O. S. (1994). Estimation of sampling variance of molecular marker data using the bootstrap procedure. *Theoretical and Applied Genetics, 89*, 259-264. https://doi.org/10.1007/BF00225151

Vos, P., Hogers, R., Bleeker, M., Reijans, M., van de Lee, T., Hornes, M, … Kuiper, M. (1995). AFLP: A New Technique for DNA Fingerprinting. *Nucleic Acids Research, 23*, 4407-4414. https://doi.org/10.1093/nar/23.21.4407

Williams, J. K. F., Kubelik, A. R., Livak, K. G., Rafalki, J. A., & Tingey, S. V. (1990). DNA polymorphisms amplified by arbitrary primers are useful as genetic markers. *Nucleic Acids Research, 18*, 6531-6535. https://doi.org/10.1093/nar/18.22.6531

Xia, Z., Tsubokura, Y., Hoshi, M., Hanawa, M., Yano, C., Okamura, K., … Harada, K (2007). An Integrated High-Density Linkage Map of Soybean with RFLP, SSR, STS, and AFLP Markers Using a Single F_2 Population. *DNA Research, 14*, 257-269. https://doi.org/10.1093/dnares/dsm027

Xu, Y., & Crouch, J. H. (2008). Marker-Assisted Selection in Plant Breeding: From Publications to Practice. *Crop Science, 48*, 391-407. https://doi.org/10.2135/cropsci2007.04.0191

Zeng, Z. B. (1993). Theoretical Basis for Separation of Multiple Linked Gene Effects in Mapping Quantitative Trait *Loci*. *Proceedings of the National Academy of Sciences, 90*(23), 10972-10976. https://doi.org/10.1073/pnas.90.23.10972

Zeng, Z. B. (1994). Precision mapping of quantitative trait *loci*. *Genetics, 136*, 1457-1468. https://doi.org/10.1093/genetics/136.4.1457