Chapter 3

Genomics in Horticultural Crops

Ali Ahmad Naz and Muhammad Jafar Jaskani♦

Abstract

Trait diversity in horticultural crops is orchestrated by a complex but orderly network of genes. Recent advances in genome and molecular analyses have made it possible to dissect and understand the hidden cues of trait diversity at molecular level. These developments offer new opportunities to improve valuable traits in crops using genomic tools. This chapter presents a brief introduction of genomics techniques and their utilities in horticultural crops.

Keywords: DNA markers, DNA sequencing, genomics, genomic resources, horticulture, linkage analysis, marker assisted selection.

3.1. Introduction

The term genome refers to entire heredity information carried by organisms. It is present in each cell in the form of long stretches of DNA **(deoxyribonucleic acid)** which instructs organisms to develop, reproduce and maintain their life cycles. These heredity instructions are laying in nucleus of the cell as nuclear or chromosomal DNA as well as in cell organelles as mitrochondrial and chloroplasts DNA. The chromosomal DNA is linear and carries most of the heredity material; whereas,

♦Ali Ahmad Naz*
Crop Genetics and Biotechnology Unit, Institute of Crop Science and Resource Conservation, University of Bonn, Katzenburgweg 5, 53115 Bonn, Germany.
*For correspondance: a.naz@uni-bonn.de

Muhammad Jafar Jaskani
Institute of Horticultural Sciences, University of Agriculture, Faisalabad, Pakistan.

Managing editors: Iqrar Ahmad Khan and Muhammad Farooq
Editors: Ahmad Sattar Khan and Khurram Ziaf
University of Agriculture, Faisalabad, Pakistan.

chloroplastic and mitochondrial DNA consist of circular plasmids like prokaryotic genome (Griffiths et al. 2002). Although, mitochondrial and chloroplast genomes bear little heredity material as compared to nuclear genome but these contain essential genes controlling fundamental cellular and subcellular processes. The inheritance of chromosomal DNA is parental (50% maternal and 50% paternal); whereas, chloroplastic and mitochondrial genomes are mostly maternal in plants.

The presence of cell organellar genome is a rule rather than exception in plants. Why plants have such genomes is one of the fundamental questions of plants evolutionary biology. A putative answer was proposed by the Russian biologist Konstantin Mereschcowsky (1855–1921) in the form of famous endosymbiotic theory (Fig. 3.1). According to this theory, smaller and less complex cells develop symbiotic relationships which appear to be the main driving force in the evolution of more complex cells. According to this, the ancestral eukaryotic cell engulfed an aerobic bacterium (proteobacteria) and a cyanobacterium in two consecutive events of cellular evolution which recruited the processes of aerobic respiration and photosynthesis in plant cells. This idea is supported by the similarities of mitochondria and chloroplasts with the aerobic bacteria and cyanobacteria, respectively (Alberts et al. 2002).

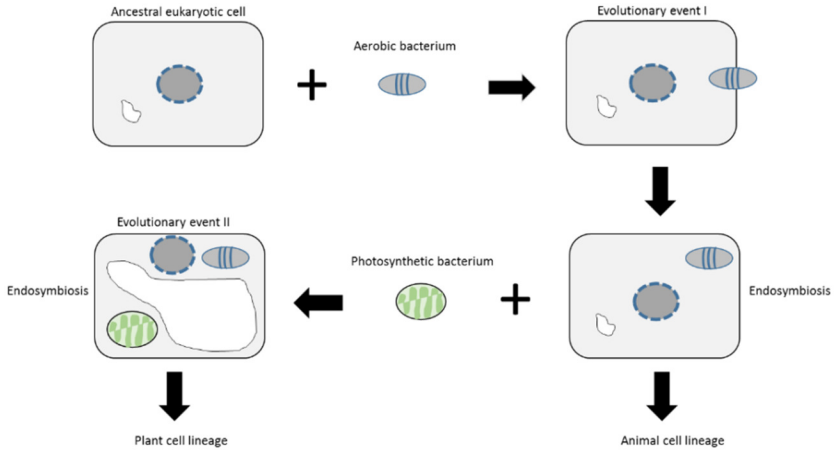


Fig. 3.1 Diagrammatic presentation of the endosymbiosis theory, according to which, complex life forms are originated from the simpler organisms. According to this theory, aerobic and photosynthetic bacteria established a symbiotic relationship inside the ancestral eukaryotic cell. By this evolutionary recruitment, the host cell transformed into a complex autotrophic plant cell, and aerobic and photosynthetic bacteria encapsulated as mitochondria and chloroplast, respectively.

The presence of DNA and RNA (ribonucleic acid) in these cell organelles is another reason to believe in this hypothesis. In the follow up investigations, the genome sequence comparison between mitochondria and proteobacteria as well as between chloroplast and cyanobacteria found significant clues of genome conservation suggesting that mitochondria and chloroplasts to be originated from aerobic bacteria

and cyanobacteria, respectively. These molecular evidences may satisfy the question, 'why these cellular organelles contain their own genomes'.

The plants show remarkable variation in their forms, shapes and agronomic traits. For instance, wild apple (*Malus sieversii*) and cultivated apple (*Malus domestica*) present dramatic difference in their fruit's size, shape and colour. Similarly, immense diversity can be seen in other horticultural crops, which expand clearly from interspecific to intergeneric levels. Recently, these morphological differences have been compared with the amount of DNA sequence difference in their genomes which clearly reveals that this diversity is proportional to the variation in their underlying genomes. Simply, the more genome variation among species of a genus, the more are the differences in their traits useful for breeding. It also suggests that important traits of horticultural crops are under the strict control of their genomes. Therefore, first-hand knowledge of their genome sequence and its analysis is crucial to understand these traits and to use them effectively in the improvement of horticultural crops.

3.2. Genomic Resources in Horticultural Crops

Since Gregor Mendel (1822-1884) performed crosses in peas and postulated the principles of inheritance, horticultural crops are the piece of cake for geneticist to understand genetics. Economically, horticulture is one of the important sectors of agriculture which deals with diverse species of fruit, vegetable, tree, ornamental, aromatic, medicinal, spices, and flowering plants. The increasing demand of horticultural products in agriculture and their fascinating genetic diversity are major driving forces behind the development of valuable genomic resources. First-hand knowledge of these resources is vital in improving production, management practices, breeding strategies as well as engineering viable horticultural crops for sustainable agriculture.

The development of the state of the art next generation sequencing has revolutionized the process of creating useful genomic information in horticultural crops. In this regard, the biggest breakthrough has been made in the whole genomic sequencing projects. In the recent past, a number genome sequencing projects have been initiated and completed in horticultural crop including fruits, vegetables, tree and ornamental plants. Among the vegetable crops, tomato (*Solanum lycopersicum*) genome is completely sequenced that deliver 950 Mb of DNA sequence. The annotation of this genome sequence has revealed 34,727 protein coding genes in tomato. Similarly, the draft assembly of potato genome sequence is also available that deliver almost similar number (35,004) genes in *Solanum tuberosum*. These data indicate a huge genetic similarity among the member of the family Solanaceae that open the doors of comparative genetics in related crops like peppers, tobacco, eggplant etc. In addition, tomato being model crop for the dicots, it is believed that its genome sequence has implications for other plant species like strawberries, apples, melons, bananas and many other fleshy fruits that share some characteristics with tomatoes. The mechanism and genes regulating fruit development is well understood in tomato, so information about the genes and pathways involved in fruit ripening, fruit shape,

size, and weight, can potentially be useful to extend them to additional fruit and vegetables crops to improve food quality and production (Tanksley 2004). An online platform (Sol genomics network) is providing value genomic resources for family Solanaceae including, whole genome assembly, gene and protein annotations, large insert genomic clones, tags for the expressed genes, expressed sequence tags (ESTs), a variety of DNA marker systems, linkage maps in selected segregating populations. Similarly, sequence assembly of citrus genome has been assembled for two citrus varieties, sweet orange and Clementine mandarin. In addition, draft sequence assemblies of papaya (*Carica papaya*), cucumber (*Cucumis sativus*), apple *(Malus x domestica*) and grape (*Vitis vinifera*) are available, thus paving the way to use the genetic potential of these crops in horticulture and to extend them further in the related crops.

The whole genome sequencing projects are delivering full length genomic and genic information in crops. Normally, such genes sequence information is more useful for the reverse genetic approach where the effect of a particular gene is traced back to the phenotypes. For this, mutant copies of these genes are needed to compare them with the respective wild-type in order to conclude the gene function in the determination of particular crop trait. Such resources are available in term of mutagenized population/ TILLING platforms (Rashid et al. 2011) or induced mutant libraries like T-DNA insertion mutant lines in Arabidopsis and to some extent in tomato, *e.g.,* ripening mutants (Osorio et al. 2011) and wolly (wo) mutant (Shilling 1959). However, the development of such resources is still the biggest challenge in horticultural crops to understand the function of genes in plant biology as well as in the expression of economically important traits, *e.g.,* fruit shape characteristics of Howard German and Banana Legs of tomato cultivars (Brewer et al. 2007; Fig. 3.2).

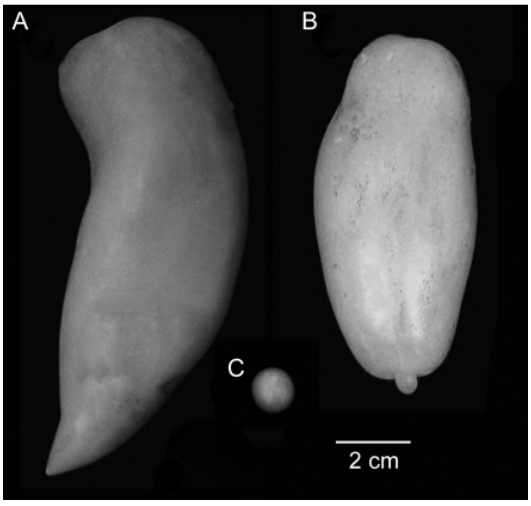


Fig. 3.2 Fruit shape and size of (A) *S. lycopersicum* cv. Howard German, (B) *S. lycopersicum* cv. Banana Legs, and (C) *S. pimpinellifolium* accession LA1589. The size bar represents 2 cm.

In this scenario, forward genetic approaches provide us solutions and luxury to dissect the natural variants of important horticultural traits. In forward genetics, any trait variation of horticultural value can be analyzed to detect its underlying gene, by making classical crossing of two variant genotypes, in the segregating populations.

Here, resources like DNA marker are essential, and will remain crucial in future for all kind of mapping projects in horticultural crops. Currently, conserved ortholog set (COS), SSR and SNP are being successfully genotyped to develop detailed genetic maps in horticultural crops like apples, citrus, tomato, potato, melons, cucumber, pepper, popular, roses etc. The development diagnostic DNA marker in these species provides an opportunity to extend these tools to additional horticultural crops via comparative genetics.

Comparative genetics is undoubtedly a powerful approach in the genetic understanding of horticultural crops. These crops revealed immense gene conservation among the related species as well as among the different species of dicots. This provides an opportunity to compare the genes among horticultural crops in order to understand their common or different genetic mechanisms. For instance, Naz et al. (2013) discovered and characterized the function of *Trifoliate* (*Tf*) gene in tomato. This gene controls the development of leaf and shoot via a conserved genetic mechanism by delaying differentiation in tomato development. The overexpression of this gene in wild-type tomato revealed the production of excessive shoots in tomato. Sequence of *Tf* gene showed significant homology across related species, suggesting a putative conservation of its molecular mechanism in dicots. These data suggest that function of this gene may be crucial in altering shoot forms in tomato and related species; therefore, *Tf* gene may be highly demanding for the cut flower crops like rose and chrysanthemum etc., where more branching is the major determinant of their production. Thus, *Tf* gene can be tested in these horticulture crops via comparative genetics approach to elucidate and extend its utility in crop production. Hence, the development of genetic resources and molecular tools are the fundaments to understand gene functions and their extension among the horticulture crops via comparative genetics. The basic knowledge of these genes is opening new doors of their utility in the horticulture industry.

3.3. Genome Structure and Its Organization

Heredity material is organized in the form of chromosomes in the nucleus of the cells. Structurally, chromosomes are composed of DNA and special structural proteins called histones. The double helix DNA is coiled on an octamere of histones molecules known as H2A, H2B, H3 and H4 (Krebs et al. 2013). An octamere histone molecule coiled with DNA double helix makes the structural unit of chromatin material called nucleosome. These nucleosomes are chained together to form the chromatin fiber inside the chromosomes. At each nucleosome, DNA double helix make two turn around the histone octamere allowing around 146 base pairs of DNA to accommodate at single nucleosome. In addition, nucleosome undergoes a special packaging protocol by which huge DNA sequences are accommodated in tiny nucleus. For example, a linear physical length of human diploid DNA is around 6 feet which follow a tight programmed packaging to fit within the nucleus. The packaging in plants DNA is no more different than human but even some plants have much more DNA than human. For example, the largest plant genome known so far, is present in *Paris japonica*, an ornamental plant that inherit 150 gigabases (Gb); whereas, human contain 3.2 Gb of DNA.

The fundamental building blocks of DNA are nucleotides, which are composed of a pentose sugar, a phosphate and a nitrogenous base (Krebs et al. 2013). The nitrogenous base can be a purine; double cyclic ring (guanine or adenine) or pyrimidine; single cyclic ring (cytosine or thymine), which are attached to the 1´ (one prime) carbon on a pentose sugar by the help of glycosidic bonds with the N_1 of pyrimidine and the N_9 of purine. A nitrogenous base attached to the pentose sugar is called a nucleoside, which is further linked to a phosphate at the 5´carbon to make a nucleotide. In addition, this phosphate linked to 3´carbon of the pentose sugar in the next nucleotide via phosphodiester bonds and by this linking nucleotides together in a polynucleotide chain. The presence of a particular nitrogenous base (guanine, adenine, cytosine, thymine) at the pentose sugar determine the identities of an individual nucleotide. As phosphate group of the 5´carbon is always attached with 3´carbon of the next nucleotide within a chain resulting in a free 5´phophate group and a free 3´hydroxyl group at the terminal nucleotides. This pattern determines 5´ to 3´ polarity in the DNA strand; whereas, its complementary strand runs in 3´ to 5´ polarity; therefore, DNA is double stranded molecule of two antiparallel strands. The nucleotides of antiparallel strands are linked with each other via hydrogen bonds, which keep two strands together in the DNA double helix. The nucleotides bonding between the strands are fixed; guanine always attaches with cytosine and adenine is linked to thymine.

The heredity information is lying in the DNA double helix in the form of genes. A variable number of nucleotides and their relative arrangement, which has the capacity to make a protein, is called a gene. Initially, the gene sequence is transcribed into RNA via a process called transcription in the nucleus. This RNA molecule move from the nucleus to cytoplasm where it gets translated into protein by ribosomes in the process called translation. These proteins are the functional units of heredity and variation in plants. However, some genes are unable to be translated into protein e.g. transfer RNAs (tRNAs), which are not translated but help in assembly of amino acids at the site of peptide chain formation for production of protein. The non-coding DNA supposed to provide additional source of genetic variation by modulating the regulatory jobs for protein coding genes. Therefore, any variation in the gene coding DNA or noncoding regulatory DNA sequences can be consequential to trait variation in horticultural crops. A major challenge of today research is to understand genetic basis of the immense trait variation in horticultural crops at DNA level to use it effectively for crop production and breeding. By the advent of modern tools and state of the art techniques in the genome analysis, it has become possible to discover the heredity messages hidden in the genome and to realize their effects on the trait performance in horticultural crops.

3.4. Methods of Genome Analysis

DNA is the heredity molecule which passes on from one to the next generation. Structurally, it consists of long chain of nucleotides; adenine, guanine, cytocine and thymine. The number and relative arrangement of these nucleotides is unique in each individual. However, different individuals of same specie or of different species revealed differences in number and arrangements of these nucleotides. These unique

differences can be used to differentiate individuals on DNA level and thus revealing them as landmarks of genetic variation. Such landmarks in DNA across the genome are known as DNA markers. Hence, DNA or molecular markers are short sequences of DNA appearing as recognizable flags on the genome comparable to the milestones on a highway (Collard et al. 2005).

DNA markers have been arisen due to different classes of DNA mutation like substitution (point mutation), re-arrangement or error in the replication of tandem repeats of DNA (Paterson 1996). Indeed, these mutations established genetic differences between individual organisms at a particular locus that is described as DNA polymorphism. Hence, a marker revealing genetic difference between two individuals of same or different species is called a polymorphic marker, and can be used as a molecular tool in plant breeding and genetics. For example, two parents, P1 and P2 showed a genetic difference (DNA sequence difference) at marker locus X. Polymorphism of both parents and their F_1 and F_2-prognies has been illustrated in a hypothetical Fig. 3.3. This marker locus can be amplified via polymerase chain reaction (PCR) that revealed the PCR product of 550 bp and 400 bp in P1 and P2, respectively. A DNA polymorphism of 150 bp exists between these two parents which can also be used to genotype their next generations like F_2 population for identifying F_2 plants bearing homozygous P1 (AA), homozygous P2 (aa) or heterozygous (Aa) alleles at marker locus X.

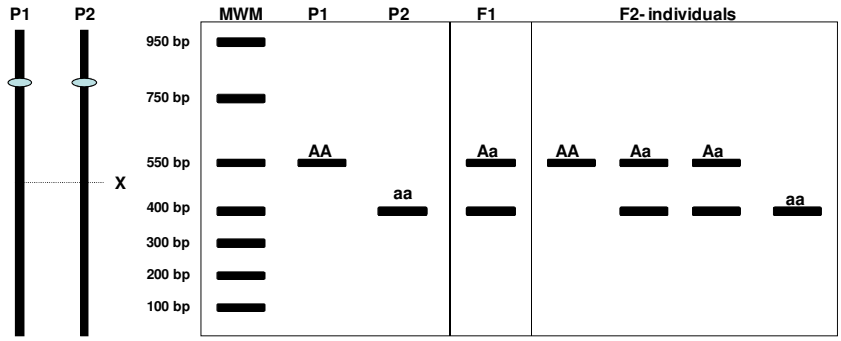


Fig. 3.3 Schematic representation of a DNA polymorphism on a electrophoretic gel between two parents, P1 (mother) and P2 (father) and among their F_1 and F_2 progenies at marker locus X. Localization of marker locus X on the chromosome (left) and genotypes of P1, P2, F_1 (first filial generation) and F_2 (second filial generation) individuals (right) after the amplification of marker locus X by PCR. AA homozygous P1, aa homozygous P2, Aa heterozygous, MWM molecular weight marker in base pairs.

A number of DNA marker types have been developed based on the kind of DNA polymorphism between two individuals and upon their method of detection across the genome.

3.4.1. Restriction Fragment Length Polymorphism (RFLP)

RFLP is the founder technique in the discovery of DNA markers and can develop DNA markers across any plant genome. It is based on the hybridization of digested DNA with the cloned DNA. Initially, organisms' DNA are fragmented with the help of enzymes called restriction endonucleases. The presence and absence of the restriction sites result in different sizes of DNA fragments. These fragmented differences are visualized by hybridizing them with their corresponding labeled DNA probes, which are synthesized by DNA cloning procedure. RFLPs can also be used to identify different individuals based on their restriction patterns differences. Such differences indicate that the restriction enzyme cut the DNA at two unrelated locations. These similarities and differences can be used to differentiate species, races and strains from one another. This marker system can be utilized for a thorough genome analysis of plant species of known genome sequences. It reveals codominant polymorphism, and therefore it remains effective for multiple genetic analysis procedures like linage mapping, phylogenetic and quantitative trait locus (QTL) analyses etc.

3.4.2. Randomly Amplified Polymorphic DNA (RAPD)

By the advent of PCR-technology, a battery of marker systems has evolved in a short period. For instance, randomly amplified polymorphic DNA (RAPD), a molecular marker based on the PCR amplification of DNA by the help of short oligonucleotide sequences (10-20 nucleotides) called random primers. It is a random amplification because the arbitrary primers bind to their complementary sequence randomly on the target DNA. In case a DNA difference exists between two individuals, these arbitrary primers might anneal or fail to anneal in one of the individuals lacking complementary sequence at the primer binding site. PCR will fail in the later individual and no PCR product is achieved. Such marker system that reveals polymorphism on the basis of presence and absence of PCR products is termed as dominant marker. RAPD does not require DNA sequence information of the targeted individuals. This marker system has been used successfully for characterization of the phylogenetic relationships of plants but it reveals lower robustness.

3.4.3. Amplified Fragment Length Polymorphism (AFLP)

AFLP is a marker system based on the combination of DNA restriction and its subsequent PCR amplification to detect restriction site polymorphisms. In first step, DNA of the target individual is digested with the combination of a rare cutter (like EcoRI) and frequent cutter (like MseI) restriction enzymes, which produce DNA fragments of optimal sizes with different sticky ends. Subsequently, the ends of these restriction fragments are ligated with specific adapters designed for EcoRI and MseI cuts. In second step, these ligated fragments are amplified using PCR primers complementary to adaptor and a part of restriction site sequence. If two genotypes differ upon their DNA sequence at particular restriction site, the respective enzymes will not cut the DNA of the individual lacking this site followed by lack of adapter-ligation and failure of PCR and results into a dominant polymorphism. It has been

considered a great marker system for developing genetic maps, DNA fingerprinting as well as for population genetics etc.

3.4.4. Simple Sequence Repeats (SSR)

SSR can be a unit of 2-6 nucleotides repeated up to 100 times at a particular locus in the genome. For example, (AT)20 is a dinucleotide consisting adenine and thymine repeated 20 times in the DNA of an individual at a particular location. Such sequences have a great potential for developing a DNA marker because they are present throughout the genome and are highly polymorphic due to high mutation rate affecting the number of repeat units during DNA replication process. DNA sequence information is needed for such marker system. Forward and reverse primers can be designed based on the sequence information and a particular SSR locus is amplified via PCR. A high resolution gel is needed to see minute polymorphism from 2 to 10 bp. However, larger polymorphisms (> 20 bp) can also be visualized on an agarose gel. It is one the most successful marker system in crop plants being highly robust, informative and multipurpose marker system.

Although, SSR genotyping is easy, being PCR based marker, but the discovery of new SSR marker in sequence deficient crops species is very critical and demands significant technical knowhow.

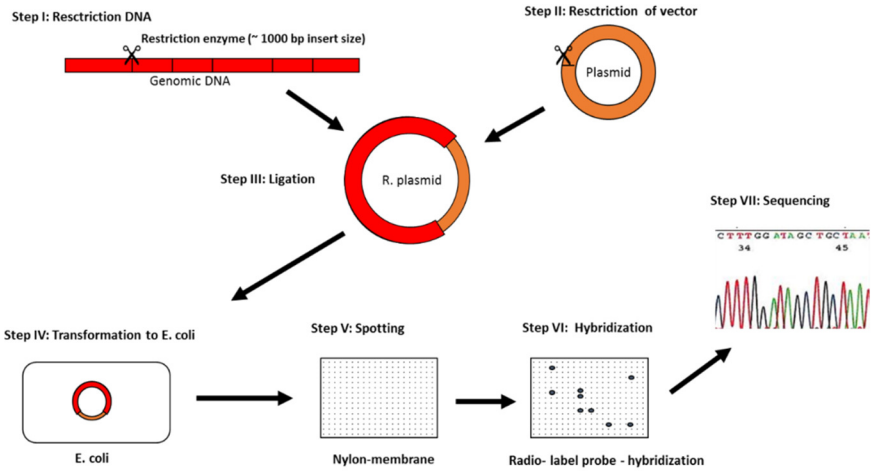


Fig. 3.4 The method of discovering new SSR markers. Firstly, the genome is fragmented by the help of appropriate restriction enzymes and ligated into a vector for cloning. Later, the target repeat units are detected among the DNA clones by hybridizing with synthesized radio-labeled probes. The positive clones are then sequenced to identify SSR across the genome.

For this, first DNA is restricted with an appropriate combination of enzymes to fragment the genome into pieces. These fragments are then ligated in a high copy vector for their cloning in *Escherichia coli* bacterium. After cloning, DNA fragments are spotted on nylon membranes for hybridization. In parallel, arbitrary radioactive DNA probes are synthesized representing individual repeats units like (AT)10. Later, these probes are hybridized with the DNA spotted membranes to detect the positive clone carrying the candidate repeat units. Subsequently, the positive clones are then sequenced using border primers of the vector to find DNA sequence of the SSR markers (Fig. 3.4).

3.4.5. Simple Sequence Length Polymorphism (SSLP)

SSLP represents DNA sequence length polymorphism like SSR, but of relatively bigger sizes. The macro mutations in the DNA like bigger insertion or deletion result into a DNA sequence length differences at a particular locus in different individuals. Thus, such locus can be amplified via PCR using primer encompassing the DNA mutations and PCR products are visualized on the gel that represents the difference of a DNA deletion or insertion. Generally, larger insertions or deletions larger than 50 bp are recommended to be genotyped via SSLP in order to visualize them easily on agarose gel. It is one of the easiest and straightforward marker systems but the abundance of such markers is low because of rare events of bigger mutations in related species.

3.4.6. Cleaved Amplified Polymorphic Sequence (CAPS)

CAPS is a method to identify the DNA polymorphism based on restriction fragment analysis of a PCR product. For this, a locus is amplified by PCR and then cleaved with an appropriate restriction enzyme. If two individuals have a DNA sequence variation that alter the restriction-site, will result into different patterns of restriction fragments. These patterns can be used to differentiate between the genotypes in term of homozygous or heterozygous alleles, thus it is a co-dominant marker system. It is very successful marker system because it is highly sensitive to all kind of crucial DNA differences, which have been arisen by insertions, deletion, rearrangement or translocations etc.

3.4.7. Single Nucleotide Polymorphism (SNP)

SNP is a technique to genotype markers based on single nucleotide differences within the DNA sequence of two or more individuals. Because of the abundance of such DNA differences across the genome, it is highly effective technique for a thorough genome analysis in crop plants. It is also highly robust method and being used in finding genetic differences of closely related species as well as within the individual of the same species or among cultivated varieties. Classically, SNP genotyping were carried out by using standard DNA sequencing approach like chain termination method (Sanger sequencing). According to this, SNP marker regions were amplified via PCR and sequenced to detect the polymorphism across genotypes.

Later, pyrosequencing, sequencing by synthesis approach, was employed for a quicker SNP detection. This technique is based on the PCR amplification of the DNA template to be sequenced together with a downstream enzyme-substrate activity to detect the incorporation of a particular nucleotide in the PCR reaction. Because the nucleotide detection is carried out during the synthesis of PCR, it is known as sequencing by synthesis. The principle of this technique is learned from the famous lightening insect, firefly (Jugnoo), which converts adenosine triphosphate (ATP) to light using luciferin in the presence of enzyme luciferase. Likewise, in this sequencing PCR, two different substrates (adenosine 5´ phosphosulfate, APS, and luciferin) and their respective catabolic enzymes, ATP sulfurylase and luciferase, are added along with the standard PCR ingredients (DNA template, primers, polymerase etc). As the first nucleotide is incorporated on the DNA template by polymerase enzyme, it releases an inorganic phosphate which is known as pyro phosphate (PPi). Later, PPi will be converted to ATP by the catabolic activity of APS and ATP sulfurylase. The resultant ATP molecule is converted to a light signal by luciferin in the presence of luciferase. This light signal produced will act as a sign of nucleotide incorporation in the growing chain of PCR. The addition of a particular nucleotide is controlled via a computer mediated micropipettes in PCR cycler. The enzyme apyrase cleans the PCR reaction after each nucleotide incorporation reaction. By the periodic incorporation of different nucleotides, signals are captured, which are later converted to DNA sequence of the template. Recently, pyrosequencing is simplified by the development of fluorescently labelled nucleotide, which has opened a new era of genome analysis called next generation sequencing, *e.g.,* 454 pyrosequencing, Illumina sequencing by synthesis etc.

3.4.8. Array Based High throughput SNP Genotyping

It is the most recent revolutionary technique to genotype thousands of SNP markers across the whole genome, simultaneously. This technique based on the amplification of allele specific PCR products and their subsequent hybridization with the locus specific DNA tag on a microarray plate (Fig. 3.5). The utmost pre-requisite of this technique is to gather short DNA sequence information (100-200 bp) across the genome that shows polymorphisms among different individuals. This sequence information is used in two folds, i) oligo (primers) synthesis to amplify thousands of SNP markers across the genome, ii) determination of a locus specific address sequence of each SNP marker and tagging of this address sequence in the microarray plate.

For each SNP, two alleles specific and one locus specific oligos are synthesized and mixed to establish an oligo pool of thousands of SNP to be genotyped in a single assay. The oligo pool is then added to DNA template where complementary primers bind to the specific SNP locus. After binding, DNA sequence between the allele specific and locus specific primers is extended and ligated with the help of appropriate enzymes. Allele specific primer binding help the synthesis of allele specific templates in combination with locus specific primer. However, in case of heterozygous SNP, both alleles have the possibility to produce templates of both alleles. On these templates, a multiplex PCR is made to amplify each SNP sequence

where fluorescently labelled dyes cy3 and cy5 are incorporated to the universal primers complementary to ASO1 (Allele Specific Oligo; allele 1) and ASO2 (allele 2), respectively. Afterwards, labelled PCR templates are separated for the hybridization to their respective address sequences on the micro array plate. For this purpose, labelled templates of single genotype are hybridized one by one and allele scoring or SNP calls are read from the micro array plates. The intensity of cy3 or cy5 or cy3-cy5 determines a SNP call for particular locus to be homozygous allele1, homozygous allele2 or heterozygous, respectively.

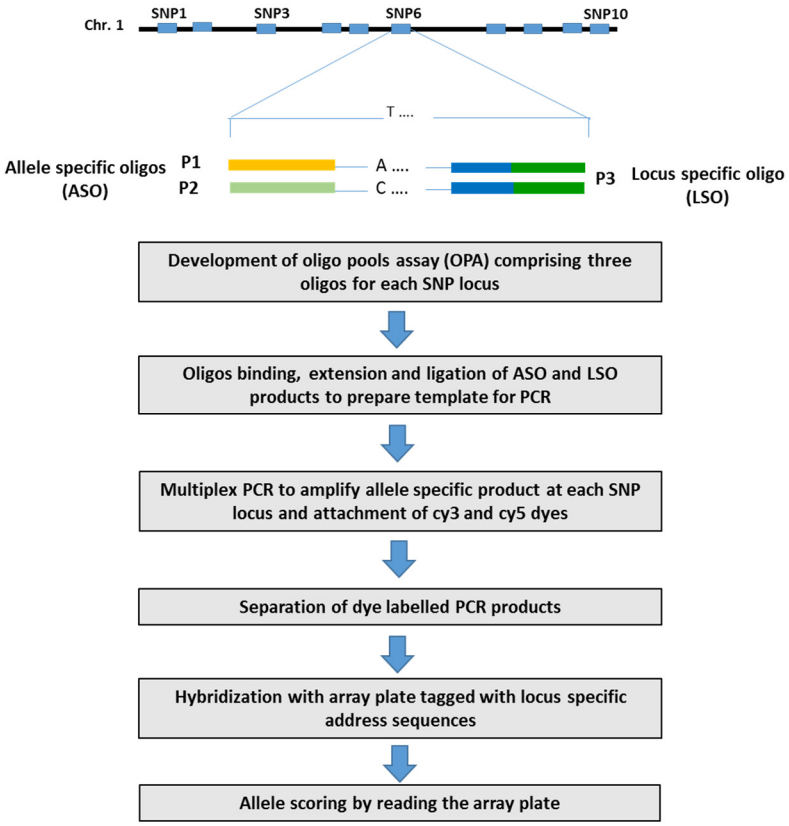


Fig. 3.5 High throughput SNP genotyping in plants. Biallelic SNP loci, representing two different alleles, amplified via the combination of allele specific and locus specific primers. During this amplification, allele 1 and allele 2 are tagged with Cy3 and Cy5 dyes. Later, the tagged fragments are hybridized with an array plate consisting of the locus specific DNA (address) sequence of individual SNP loci. After hybridization, allele genotyping is done by reading the array plate where Cy3, Cy5 and Cy3-Cy5 are scored as homozygous allele 1, homozygous allele 2 and heterozygous, respectively (according to habilitation work of author "Ali Ahmad Naz").

3.5. Application of DNA Marker in Plant Breeding

Molecular markers are useful tools for developing detailed linkage maps that are necessary for different kinds of genetic analyses. A linkage map determines how two markers loci are connected with each other and assign their positions on a chromosome. These maps are essential for genetic mapping, QTL analyses, and marker assisted selection as well as for the positional cloning of genes.

3.5.1. Genetic Linkage Analysis

During meiosis, a chromosome exchanges its segments with its homologous pair by a phenomenon called genetic recombination. By this, different gametes are formed with new allele combinations. This process is the major source of variation in plants traits.

Normally, two adjacent recombination events happen on chromosome at a considerable distance because of the phenomenon of genetic interference and therefore, there is a high chance that marker loci that are inherited together, lying physically closer and therefore tend to stay together in next generation. This rare occurrence of recombination events is called as genetic linkage and it is measured in centiMorgans (cM). It is also known as genetic distance or recombination frequency. Two marker loci are said to be 1 cM apart, if 1% of the individuals recombine at these two markers. In tomato, a genetic distance of 1 cM may correspond to around 300 to 500 kb on the physical length of the chromosome. However, it varies very much with respect to different regions on the chromosome and between different species. The higher recombination frequency between two markers, the higher will be the genetic distance resulting in a lower genetic linkage and vice versa. A relationship of recombination and cM distance is shown in Fig. 3.6. M1 and M2 are 1 cM apart from each other because a single recombination event happened between these two markers among 100 meiotic events. Likewise, 4 crossing over events happened between M2 and M3 and therefore these are 4 cM apart from each other. Hence, M1 and M3 are at the 5 cM genetic distance. This relationship seems additive which holds true only in an ideal case of complete genetic interference across the chromosome or within small distance on the chromosome. In this case, the recombination fractions are proportional to the genetic distance (genetic distance, d = recombination fractions, r). However, genetic interference does not exist always during crossing overs and quite often double recombination events happen that can disturb the relationship of d and r. In order to eliminate this discrepancy, two different kinds of mapping functions are used depending upon the level of genetic inference existing among the recombination breakpoints across the chromosome. The Haldane mapping function is appropriate for the chromosomal regions or chromosome with no interference; whereas, the Kosambi mapping function allows some interference which normally happens during genetic recombination. Two markers are called genetically unlinked if the genetic distance between them is around 50 cM or more.

Although linkage groups can be calculated manually for few markers but it is not feasible to manually calculate the recombination frequencies of large number of

markers. These days, computer program are available to calculate the genetic distances between markers and simultaneously drafting a linkage map. Genetic linkage between markers is usually calculated using odds ratios, the ratio of linkage versus no linkage. It is called as logarithms of odds (LOD) value or LOD score (Risch 1992). LOD values of > 3 are typically used to construct linkage maps. A LOD value of 3 between two markers indicates that linkage is 1000 times more likely than no linkage. Commonly used software programs include Mapmaker/ EXP (Lander et al. 1987; Lincoln et al. 1993) and MapManager QTX (Manly et al. 2001).

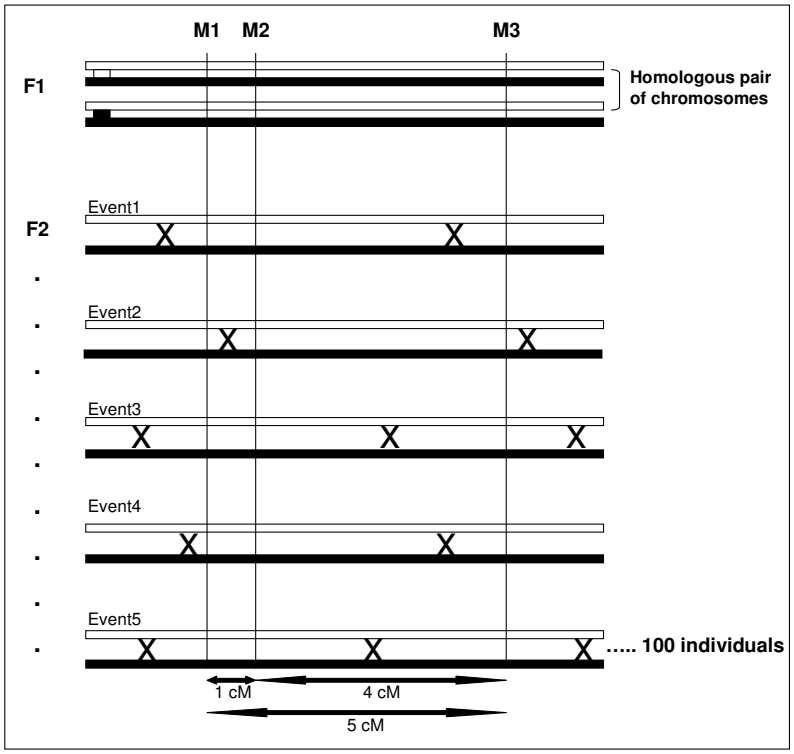


Fig. 3.6 Hypothetical representation of link between the crossing over events and genetic distance. The number of crossing over events is a measure of the genetic distance between markers; M1, M2 and M3. The plants showing recombination break points between M1, M2 and M3 are shown in this Fig. among the 100 F_2 plants.

3.5.2. Genetic Mapping

Crop traits are determined by the action of a single or multiple genes. These genes are small entities hidden in the genome similar to a needle in the haystack. The process of identifying the linkage between traits and their controlling genes, is called genetic mapping. If a trait has a single gene inheritance, this procedure is known as gene mapping, and in case of quantitative trait, QTL mapping. The basic idea behind this technique is to divide the whole genome into smaller parts by the development

of representative DNA markers. These DNA markers will act as reference points for the nearby linked genes. Normally, this process is carried out in a segregating population like F_2 of a cross between two contrasting parents (P1 × P2) that show variation in a trait of interest. For initial mapping, the trait of interest is measured in a population size of around 200 F_2 plants. In parallel, these F_2 plants are genotyped using DNA markers to dissect the presence of alleles; homozygous alleles P1 (group 1) or homozygous alleles P2 (group2) or heterozygous alleles from P1 and P2 (group 3) at a particular marker locus. By this, three genotypic groups are established for each locus across the genome. For example, if marker X is genotyped in 200 plants, the expected allele frequencies are homozygous P1 (50 F_2 plants, 25%), homozygous P2 (50 F_2 plants, 25%) and heterozygous (100 F_2 plant, 50%). Ultimately, the significant difference of the measured traits is tested among these genotypic groups by variance analysis (ANOVA, Tanksley 1993). A significant difference between phenotypic means of these genotypic groups suggests that the marker locus X is linked to the trait of interest. Conversely, an unlinked marker does not show significant trait difference among the genotypic groups. Likewise, each DNA marker across the genome is tested one by one for their association with the measured trait; this approach of genetic mapping is also known as single point analysis (Fig. 3.7).

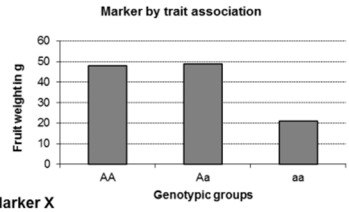
Gene and trait variation are in the relationship of a cause and the consequence. However, sometime, multiple causes resulted into a single consequence. Similarly, a polygenic inheritance implies when more genes, gene interactions or gene by environment interactions are concerted together in the expression of single trait. Such trait, which is controlled by the cumulative response of many genes, is known as quantitative or polygenic trait. If the trait under study has polygenic inheritance, then the associated marker locus is termed as QTL, and procedure of identifying linked marker to the quantitative traits is known as QTL mapping or QTL analysis. Most of the traits of horticultural importance are quantitative like fruit weight in tomato. The procedure of genetic mapping of quantitative traits is not different than standard genetic mapping, but rather more statistical depending upon the complexity of the trait. Further, more DNA marker and bigger mapping population are needed for the higher resolution QTL mapping. A summary of this process is highlighted in Fig. 3.7, in which the linkage of marker X and marker Y has been calculated using 16 F_2-plants. However, a bigger population size is needed in the real mapping experiments.

Frary et al. (2000) discovered and cloned the major QTL (*fw2.2*) controlling fruit in tomato. This QTL underlie single gene ORFX, which is expressed early in floral development and controls carpel cell number. Fruit size is one of the desiring traits during plants domestication. Similarly, *fw2.2* was selected in the cultivated varieties during tomato domestication. By virtue of this QTL effect, most of the cultivated varieties of *Solanum lycopersicum* carry giant fruits as compared to their wild relative *Solanum pimpinellifolium*.

A

F2 plants	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Marker X	aa	Aa	aa	AA	Aa	Aa	AA	Aa	Aa	aa	AA	Aa	aa	AA	Aa	Aa
Fruit weight (in grams)	22	47	25	46	51	53	50	49	47	18	49	51	18	47	46	49
Marker Y	AA	aa	aa	Aa	Aa	AA	AA	Aa	aa	Aa	Aa	Aa	Aa	aa	AA	Aa
Fruit weight (in grams)	40	35	20	20	14	24	20	30	33	50	40	25	30	25	19	19

B



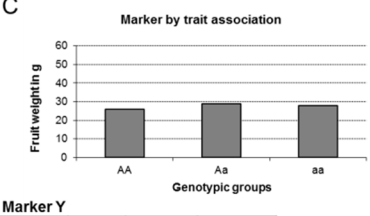
Marker X

Groups	Count	Average	Variance
AA	4	48	3.3
Aa	8	49	5.8
aa	4	21	5.7

P value= < 0.0001
Highly significant

* Marker is linked to fruit weight

C



Marker Y

Groups	Count	Average	Variance
AA	4	26	94.9
Aa	8	29	140.6
aa	4	28	48.9

P value= 0,91
Non-significant

* Marker is unlinked to furit weight

Fig. 3.7 Procedure of QTL mapping by single point analysis using genotypic and phenotypic data. A). Genotypes and phenotypes of 16 F_2-plants at marker locus X and Y. B). Analysis of variance in order to calculate the association of marker X to fruit weight. The highly significant difference in fruit weight between AA, Aa and aa showed that marker X is linked to a dominant QTL. C). A similar analysis as in B for the marker Y. Variance analysis revealed a non-significant P value and therefore, marker Y is un-linked to fruit weight.

3.5.3. Marker Assisted Selection

In the classical breeding, breeders select a required phenotype and test it reproducibility across several years in order to establish a cultivar of desired trait. It is also possible that other environmental factors interact with the phenotype and cause a great experimental error during this selection. As we know that, a DNA marker is a short piece of DNA that remains constant generation after generation and segregates by the Mendelian inheritance. By the application of DNA markers, a gene of interest can be identified and selected at a seed or seedling level. This method will also eliminate the confusion about the co-localization of the targeted gene and its phenotype. Currently, such selections are very common in plant breeding and this procedure is known as marker assisted selections. Bacterial spot (BS4) resistance gene has been discovered and utilized successfully in tomato resistance breeding through marker assisted selection (Schornack et al. 2004). Similarly, frost tolerance loci were identified from wild potato species which were successfully transferred to cultivated potato varieties through marker assisted backcrossing (Vega et al. 2003). In citrus rootstock breeding, major locus Tyr1 that confers resistance against citrus nematodes is being utilized via marker assisted selection (Xu et al. 2010).

3.5.4. DNA Fingerprinting

DNA fingerprinting is a general term that refers to the identification of specific individuals based on their DNA sequence. Molecular markers are the major tools being used successfully in this technique. The principal based on RFLP or PCR analyses of molecular marker like SSR in order to reveal the specific DNA profile for particular organism that has unique DNA fingerprints. A DNA fingerprint is generally independent of environment, and is consistent throughout different parts and developmental stages of the plants. DNA fingerprinting helps to estimate the genetic relatedness of two species as well as distinguish plants from different families, genera, species, cultivars and even sibling plants. Recently, genetic relatedness of horticultural plants and their taxonomic classifications are carried out by their DNA fingerprinting. Zhang et al. (2014) utilized genomic SSR markers for the identification and classification of Chrysanthemum cultivars in China. Currently, an innovative project, known as 'Barcode of Life' is working immensely on species identification, classification and precise differentiation of ornamental, fruit, vegetables as well as medicinal plants throughout the world (http://www.Barcodeoflife.org/). This project is focusing chloroplast and mitochondrial genome based DNA markers for species fingerprinting among many horticultural crops.

3.6. Acknowledgements

We dedicate this humble effort to the soul of senior author father Rai Bakhtawar Ali. By virtue of his inspiration, he got fascinated in education and science. A special thanks to Miss. Annemarie Bungartz for reading and editing this piece of writing, to Mrs. Esther van der Knaap and Society for Experimental Biology for the contribution of Fig. 3.2 in this chapter.

References

Alberts, B, A. Johnson, J. Lewis (2002). Molecular Biology of the Cell, 4th Edition. Garland Science, NY, USA.

Brewer, M.T., J.B. Moyseenko, A.J. Monforte and E. van der Knaap (2007). Morphological variation in tomato: a comprehensive study of quantitative trait loci controlling fruit shape and development. J. Exp. Bot. 58: 1339-1349.

Collard, B.C., Y. M.Z.Z. Jahufer, J.B. Brouwer and E.C.K. Pang (2005). An introduction to markers, quantitative trait loci (QTL) mapping and marker assisted selection for crop improvement: The basic concepts. Euphytica 142: 169-196.

Frary, A., T.C. Nesbitt, A. Frary, S. Grandillo, E. van der Knaap, B. Cong, J. Liu, J. Meller, R. Elber, K.B. Alpert and S.D. Tanksley (2000). fw2.2: A quantitative trait locus key to the evolution of tomato fruit size. Science 289: 85-88.

Griffiths, A.J.F., W.M. Gelbart, R.C. Lewontin and J.H. Miller (2002). Modern Genetic Analysis. Integrating Genes and Genomes, 2nd edition. W.H. Freeman and Co., NY, USA.

Krebs, J., E. Goldstein, S. Kilpartick (2013). Lewins´s Genes X, 11th Edition. Jones and Bartlett Publishers, London, UK.

Lander, E.S., P.G.J. Abrahamson, A. Barlow, M.J. Daly, S.E. Lincoln and L. Newburg (1987). Mapmaker an interactive computer package for constructing primary genetic linkage maps of experimental and natural populations. Genomics 1: 174-181.

Lincoln, S., M. Daly and E. Lander (1993). Mapping genes controlling quantitative traits using MAPMAKER/QTL. Version 1.1. Whitehead Institute for Biomedical Research Technical Report. 2nd Edition. Whitehead Institute for Biomedical Research/MIT Center for Genome Res., Cambridge, MA, USA.

Manly, K.F., H.C. Robert and J.M. Meer (2001). Map Manager QTX, cross platform software for genetic mapping. Mamm. Genome. 12: 930-932.

Naz, A.A., S. Raman, C. Martinez, N. Sinha, G. Schmitz and K. Theres (2013). Trifoliate encodes an R2R3 MYB transcription factor that modulates leaf and shoot architecture in tomato. Proc. Nat. Acad. Sci. USA. 99: 1064-1069.

Osorio, S., R. Alba, C.M.B. Damasceno, G. Lopez-Casado, M. Lohse, M.I. Zanor, T. Tohge, B. Usadel, J.K.C. Rose, Z. Fei, J.J. Giovannoni and A.R. Fernie (2011). Systems biology of tomato fruit development: combined transcript, protein and metabolite analysis of tomato transcription factor (nor, rin) and ethylene receptor (Nr) mutants reveal novel regulatory interactions. Plant Physiol. 157: 405-425.

Paterson, A.H. 1996. Making genetic maps. In: Paterson, A.H. (ed.). Genome Mapping in Plants. Academic Press, Austin, Texas, USA. pp. 23-39.

Rashid, M., G. He, Y. Guanxiao and K. Ziaf (2011). Relevance of tilling in plant genomics. Aust. J. Crop Sci. 5: 411-420.

Risch, N. 1992. Genetic linkage: Interpreting LOD scores. Science. 255: 803–804.

Schornack, S., A. Ballvora, D. Gurlebeck, J. Peart, M. Ganal, B. Baker, U. Bonas and T. Lahaye (2004). The tomato resistance protein Bs4 is a predicted non-nuclear TIR-NB-LRR protein that mediates defense responses to severely truncated derivatives of AvrBs4 and over expressed AvrBs3. Plant J. 37:46-60.

Shilling, P.R. (1959). An investigation of the hereditary character, Wooly, in the tomato. Ohio J. Sci. 59: 289-302.

Tanksley, S.D. (1993). Mapping polygenes. Ann. Rev. Genet. 27: 205-233.

Tanksley, S.D. (2004). The genetic, developmental, and molecular bases of fruit size and shape variation in tomato. Plant Cell 16: S181-S18.

Vega, S.E., A.H. del Rio, G. Jung, J.B. Bamberg and J.P. Palta (2003). Marker-assisted genetic analysis of non-acclimated freezing tolerance and cold acclimation capacity in a backcross *Solanum* population. Am. J. Potato Res. 80: 359-369.

Xu, X., D. Zhan-Ao, C. Chun-Xian, F.G. Gmitter Jr. and K. Bowman (2010). Marker assisted selection in citrus rootstock breeding based on a major gene locus 'Tyr1' controlling citrus nematode resistance. Agri. Sci. China 9:557-567.

Zhang, Y., S. Dai, Y. Hong and X. Song (2014). Application of genomic SSR locus polymorphisms on the identification and classification of chrysanthemum cultivars in China. PLoS ONE. 9(8): e104856.