Chapter 2
Methods for Developing Molecular Markers

Hee-Bum Yang, Won-Hee Kang, Seok-Hyeon Nahm, and Byoung-Cheorl Kang

Abstract Molecular markers are essential for breeding major crops today and many molecular marker techniques have been developed. DNA markers are now the most commonly used. This chapter describes the principles of DNA marker techniques and methods to map major genes. DNA markers can be classified into two categories: (1) DNA hybridization-based techniques, including restriction fragment polymorphism and DNA chips, and (2) polymerase chain reaction techniques, including simple sequence repeats, random amplified polymorphic DNA, amplified fragment length polymorphism, and single nucleotide polymorphism. To develop trait-linked markers, segregating populations for the target traits and reliable phenotyping methods are indispensable. With these tools, two approaches can be used to develop trait-linked markers: (1) when there is no biological information for the trait, and (2) when biological information is available. Finally, we describe several case studies for trait-linked marker development.

2.1 Definition of Technology and Related Terminology

When a specific phenotype, such as disease resistance or crop quality in plants, is difficult to determine, a different method must be used to investigate the trait. Genetic markers are a viable alternative. Because they are located close to the target gene and are inherited with it, selecting plants with useful traits using genetic markers is relatively easy. Genetic markers can be classified into morphological markers, including plant shape and/or color; protein markers, such as isozymes; and DNA markers based on sequence differences.

Author contributed equally with all other contributors.

H.-B. Yang • W.-H. Kang • B.-C. Kang (✉)
Department of Plant Science, Seoul National University, Seoul, Republic of Korea
e-mail: yhbk0130@snu.ac.kr; hui81@snu.ac.kr; bk54@snu.ac.kr

S.-H. Nahm
NongWoo Bio Co., LTD, Yeoju, Republic of Korea
e-mail: shnahm1@nongwoobio.co.kr

H.-J. Koh et al. (eds.), *Current Technologies in Plant Molecular Breeding*
DOI 10.1007/978-94-017-9996-6_2

Morphological markers were used first for genetic analysis by geneticists like Gregor Mendel and Thomas Hunt Morgan. However, their potential numbers are low, so few examples of their practical use exist. Protein markers such as isozymes, which were developed later, can distinguish individual plants. Thanks to this method, many samples could be analyzed with low cost. However, the relatively small number of isozyme variants limits their utility. DNA-based restriction fragment length polymorphism (RFLP) markers were developed and used in the 1980s, and in the following decade, polymerase chain reaction (PCR) technology gave rise to various types of DNA markers. The strengths and weaknesses of each type of genetic marker can be seen in Table 2.1.

Typically, the term 'molecular markers' indicates technology that uses phenotype-determining or closely related genes to find similarities or differences among individual plants, cultivars, or breeding lines. By analyzing molecular markers, individual plants with useful phenotypes can be selected at the seedling stage. Molecular markers for plant breeding are based on genetic differences among individuals in alleles at a certain locus. A molecular linkage map can be constructed by investigating the marker genotypes of each individual plant and calculating the genetic distances between marker pairs based on the recombinant frequency.

Table 2.1 Comparison of different types of genetic markers

Type	Benefit	Drawback	Example
Morphological markers	Easy to assay	Highly dependent on environmental factors	Color, shape, etc.
	Low cost	Difficult to analyze for quantitative traits	
		Difficult to determine heterozygosity	
		Limited availability	
Protein markers	Low cost	Assay samples must be in good condition	Isozymes
	Co-dominant	Limited availability	
	Less dependent on environmental factors	Unstable materials (protein)	
DNA markers based on hybridization	Do not require sequence information of the target	Costly and time consuming	RFLP
	Co-dominant	Use isotopes	
	Unaffected by environmental factors	Require large quantities of high molecular weight DNA	
		Difficult to automate	
DNA markers based on PCR	Require low quantities of DNA	Require expensive equipment	SSR
	Quick and easy to assay	Require sequence information	AFLP
	High accuracy		RAPD
	Unaffected by environmental factors		SNP

2.1.1 Types of Molecular Markers

2.1.1.1 Protein Markers

Proteins are the products of gene expression. Different alleles encode different amino acid sequences, giving proteins different sizes or biochemical characteristics that can be observed by electrophoresis and thus used as molecular markers. Isozymes are an example. Isozymes are useful molecular markers because they can be distinguished from each other based on differences in charge or size, despite having the same enzymatic activity. Although proteins cannot usually be seen by the naked eye, isozymes can be easily detected by separating via gel electrophoresis then adding the substrate of the enzyme. The isozyme produces color from the substrate, producing a band on the gel. Isozyme markers have a few drawbacks that greatly reduce their utility. In addition to having a very limited number of possible markers (only a few dozen have been developed), they are not distributed evenly on the chromosome, and often the enzyme activity depends on the plant's age or tissue type. Even so, isozyme analysis is very cheap and simple and was used for studies of maize, wheat, and barley decades before DNA markers were developed.

2.1.1.2 DNA Markers

DNA marker techniques use sequence differences among species or individuals within a species. Genetic differences among individuals in a group are usually due to abnormal pairing of sister chromosomes or recombination that rearranges the chromosomes, for example, insertions, deletions, inversions, translocation events, or reduplication. Chromosomal rearrangements can vary in size, from just a few base pairs to millions. DNA mutations in the form of base pair substitutions also occur. To develop genetic markers using DNA variants, DNA hybridization or PCR techniques are often used. In DNA hybridization, a short DNA fragment that is homologous to the target DNA is used as a probe. The probe is tagged with a radioisotope and hybridized with the DNA being analyzed. DNA variations can be detected based on the target–probe hybridization or the size of the hybridized DNA fragment. RFLP is an example. PCR techniques require only a small amount of DNA and are relatively simple and inexpensive. They include using minisatellites or microsatellites, sequence-specific primers such as sequence tagged sites (STSs) or expressed sequence tags (ESTs), and random primers such as random amplified polymorphic DNA (RAPD) or amplified fragment length polymorphism (AFLP) to amplify the DNA fragment and analyze its variants.

Molecular Markers Using Hybridization Methods: RFLP The classic example of molecular markers using DNA hybridization, RLFPs are a first-generation technique and the basis of many DNA marker methods that are used today. To better understand RFLPs, restriction enzymes and Southern blotting must be clear (Fig. 2.1).

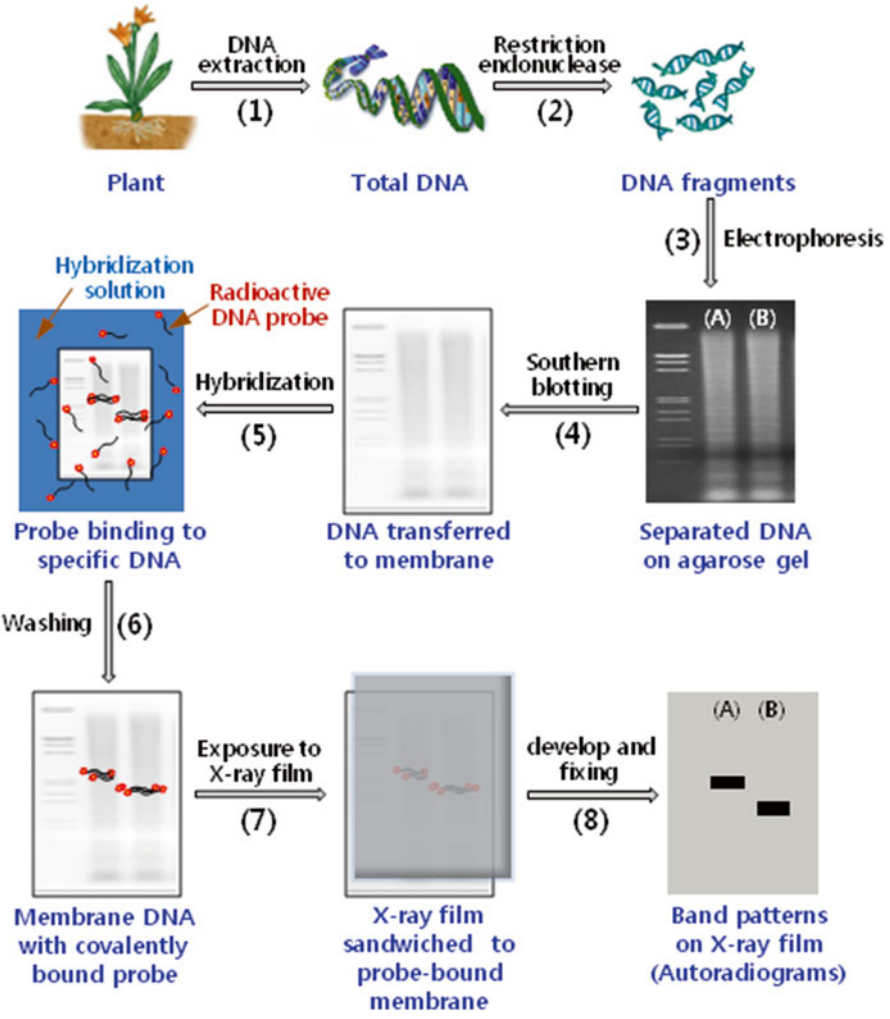

Fig. 2.1 Restriction fragment length polymorphism (RFLP) procedure. (*1*) Extract DNA from individuals A and B. (*2*) Use restriction enzymes to cut DNA. (*3*) Electrophorese DNA fragments on agarose gel to separate them by size. (*4*) Transfer the DNA in the gel to a nylon membrane by Southern blot. (*5*) Use radioactively labeled DNA fragments as probes to hybridize to the DNA. (*6*) Remove non-specifically bound or unbound probes by washing the nylon membrane. (*7*) Expose the washed membrane to X-ray film. (*8*) Develop the X-ray film to observe DNA polymorphisms

Restriction Endonucleases Restriction endonucleases (or restriction enzymes) are enzymes that recognize a specific DNA sequence (restriction sequence) and cut the DNA there. The recognized sequence can be four, six, or eight base pairs in length, depending on the enzyme. If a given DNA sequence has an equal numbers of A, T, G, and C, a restriction enzyme that recognizes six base pairs will cut every 4,096(4^6) positions on average, and the DNA of an organism with a genome size of 10^9 bp

would be cut by such a restriction enzyme into about 250,000 fragments of different sizes. If a mutation happens at the restriction sequence, the restriction enzyme will no longer cut there. RFLP technology capitalizes on point mutations, which are common and widespread. Thus, individual genotypes, even within a species, result in different patterns of fragmentation. However, when too many fragments are generated, they cannot be differentiated by typical electrophoresis. Even when they can be visualized, identifying homologous fragments in different individuals is nearly impossible. Southern blotting analysis was therefore developed to overcome these problems.

Southern Blot Analysis Southern blot analysis uses 0.5–3.0 kb DNA fragments as probes to detect other fragments with the same sequence among the enzyme-digested products. The probe is selected from a genomic DNA library, a cDNA library, or DNA fragments related to the gene(s) of interest. Southern blot analysis proceeds as follows:

1. DNA is extracted from the plant materials to be surveyed and appropriate restriction enzymes are selected.
2. DNA fragments are digested with restriction enzymes and separated by size using agarose gel electrophoresis. Fragment lengths will vary from a few hundred base pairs to over 20 kb.
3. After electrophoresis, the gel is treated in a strong alkali solution to denature the DNA strands. Then, the fragments are transferred to a nitrocellulose or nylon membrane and exposed to UV light or heat to fix them onto the membrane.
4. The DNA probe is labeled with radioactive isotope or fluorescent dye and hybridized to the membrane-fixed DNA fragments. During hybridization, labeled DNA probes will bond to complementary target fragments.
5. Finally, the membrane is washed to remove unbonded or weakly bonded probes. The washed membrane is exposed to X-ray film to visualize the hybridization band patterns.

If the restriction sites differ among individuals because of mutations, the probe banding pattern will reveal this variation (Fig. 2.2). RFLP markers have many advantages, such as being mostly codominantly inherited, highly reproducible, and distributed evenly on the genome. However, because the process is complicated and requires a large amount of pure DNA, RFLP markers are not often used today.

Molecular Markers Using Polymerase Chain Reaction (PCR) Several methods have been developed based on PCR and are summarized here.

Minisatellite/Microsatellite Markers Minisatellites are 9–100 bp DNA sequences that are repeated along the genome. The number of repeats varies but is usually under 1000. Polymorphism in the number of repeats is called variable number of tandem repeats (VNTR). VNTR can be measured by either probe hybridization or PCR. When a probe hybridizes with a fragment containing minisatellite repeats, length differences in the fragments indicate VNTR. Alternately, PCR using primers that flank the repeat allow size variation of amplicon, and thus VNTR, to be detected.

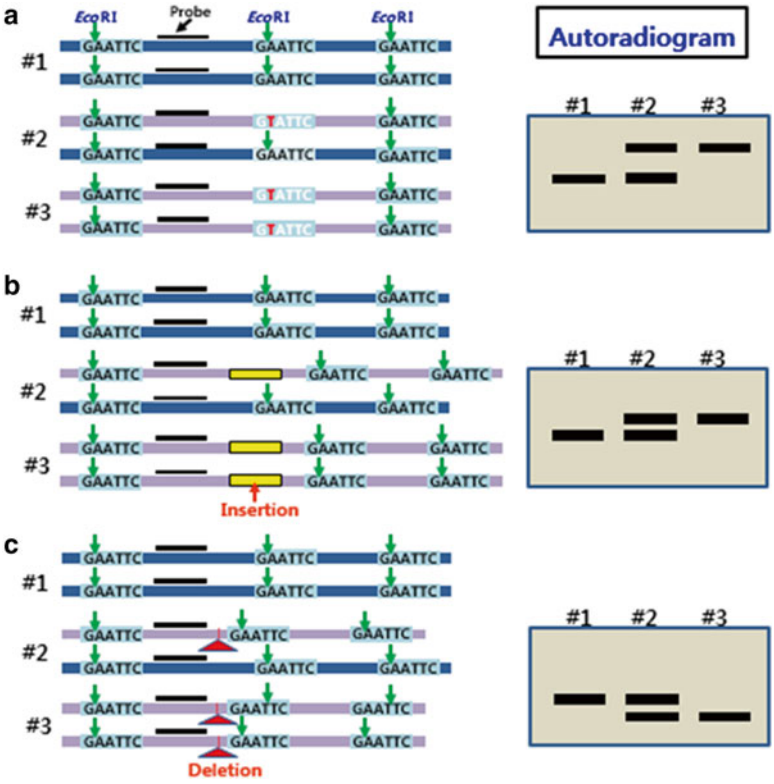

Fig. 2.2 Examples of restriction fragment length polymorphism (RFLP) analysis. (**a**) RFLP caused by substitution. A point mutation occurred at the restriction enzyme recognition site; the length of the fragment to which the probe hybridizes changes. (**b**) RFLP caused by insertion. (**c**) RFLP caused by deletion

Microsatellites, also called simple sequence repeats (SSRs), are 1–6-bp random sequences that are repeated fewer than 100 times. The most common case is two alternating bases, such as $(CA)_n$:$(GT)_n$, $(CG)_n$:$(GC)_n$, or $(AT)_n$:$(TA)_n$. Three or four base-pair repeats also exist, but are less common. In plants, the number of microsatellite repeats can differ among individuals and species. When a microsatellite repeat is discovered, the flanking sequences can be used as PCR primers to develop SSR markers (Fig. 2.3). In the past, to develop an SSR marker, a probe containing a repeat sequence was first used to identify homolog repeats in genes from a DNA library. However, because large amounts of sequence data are now easy to obtain, the need to screen a DNA library no longer exists. SSR markers can now be developed from gene bank data. After a repeat is identified, the flanking sequence is used to amplify the SSR, and length differences among the amplicons reveal SSR polymorphisms. SSR markers are usually codominant, easy to analyze, and provide reproducible results, making them ideal molecular markers.

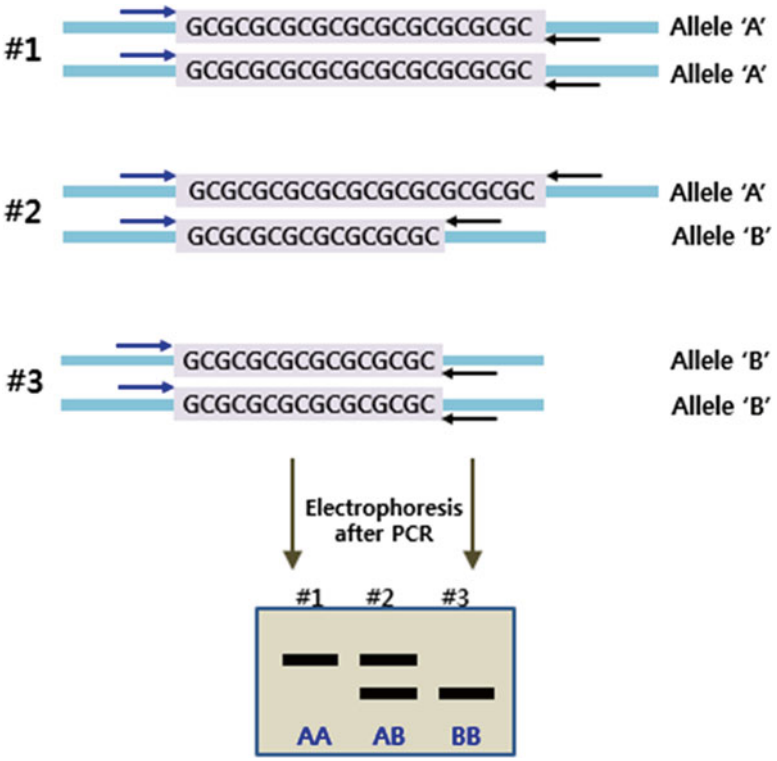

Fig. 2.3 Principle of simple sequence repeat (SSR) markers. Amplifying the SSR sequence by PCR using primers just outside the sequence reveals polymorphisms in amplicon size

Random Amplified Polymorphic DNA (RAPD) Markers RAPD markers can be developed even without DNA sequence data of the target species. These markers are generated using random PCR primers of about 10 bp in length. Because the random primers can bind with any complementary sites in the genome, when two primers are close enough to amplify the intervening sequence, the PCR product can be analyzed by electrophoresis. DNA sequence differences among individuals at the priming sites cause different band patterns, allowing polymorphisms to be observed on an agarose gel (Fig. 2.4). RAPD markers are relatively easy to develop, and their results are relatively simple to analyze. However, because RAPD markers are inherited dominantly and experimental reproducibility can vary with reaction conditions, precise genotype analysis is difficult. Despite these shortcomings, RAPD is widely used to develop markers linked to certain traits, such as disease resistance.

Amplified Fragment Length Polymorphism (AFLP) Markers To solve the reproducibility problems of RADP, the Dutch biotech company KeyGene developed the AFLP method in 1995. AFLP has the strengths of both RFLP and RAPD. In AFLP, restriction-digested DNA fragments are amplified using PCR. The basic principle of

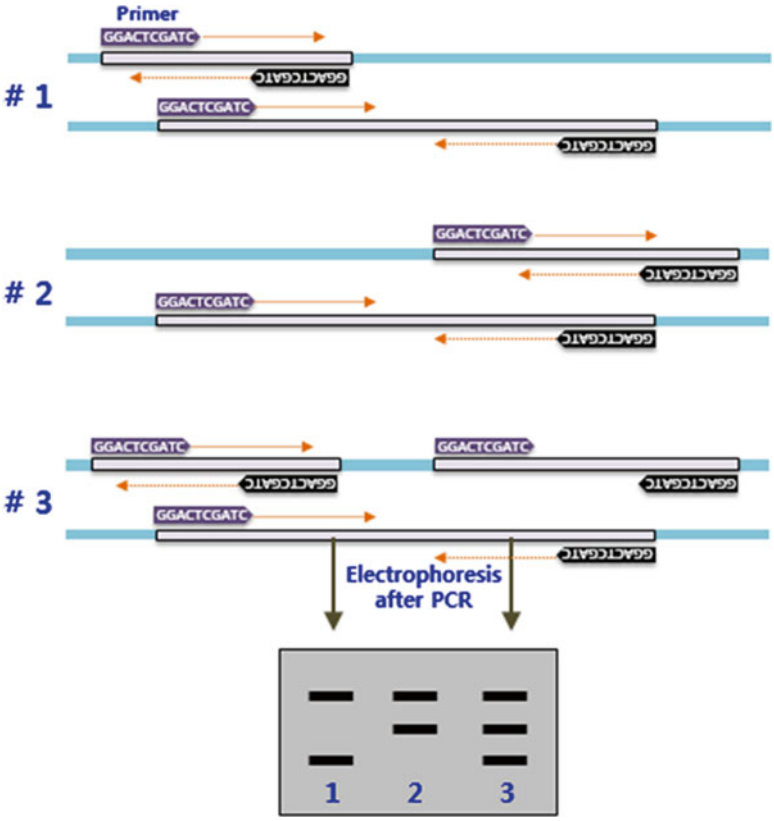

Fig. 2.4 Principle of random amplified polymorphic DNA (RAPD). A single random primer binds to complementary sites in the genome. Difference RAPD profiles on an agarose gel reveals polymorphism among individuals

AFLP is shown in Fig. 2.5. First, DNA is digested by two different restriction enzymes. An adaptor, a small double stranded oligonucleotide that complements the sticky end of the digest, is attached the end of each digested fragment. Then, the adaptor-attached DNA fragments are amplified by PCR using primers that complement the adaptors. One primer must be labeled with a radioactive isotope or fluorescence to visualize the amplified products. A few additional base pairs (selective nucleotides) can be added to the primer sequence beyond the adaptor region to decrease the number of amplified fragments. For example, if three selective nucleotides are added to each primer, the number of amplified fragments will be reduced by $1/4^6$. Changing the selective nucleotides allows different DNA fragments to be amplified.

AFLP can be used to detect differences in restriction enzyme sites, selective nucleotide sequences, and insertions/deletions among individuals. AFLP is less influenced by PCR conditions than other methods, so reproducibility is very high. Moreover, because 50–100 DNA fragments can be analyzed at once, many loci can

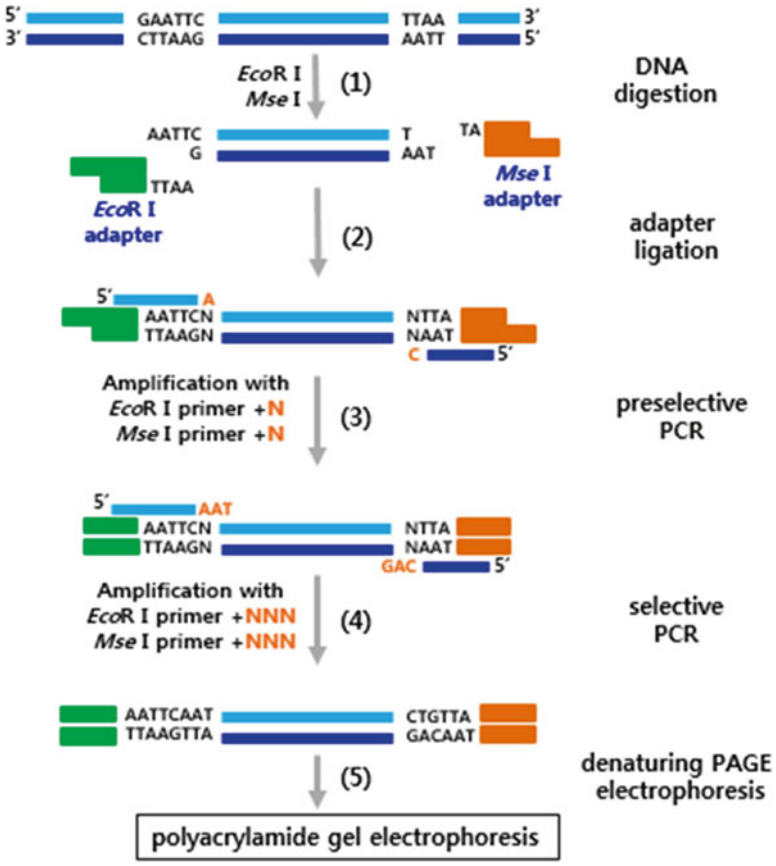

Fig. 2.5 Schematic diagram showing the amplified fragment length polymorphism (AFLP) technique. (*1*) The DNA sample is digested with two types of restriction enzymes. (*2*) Adaptors are attached to each side of DNA fragment. (*3*) Pre-amplification by PCR using primers that have the adaptor sequence and one random additional base (+1 selective nucleotide). (*4*) Selective amplification with primers that have the adaptor sequence and +2–4 selective nucleotides. (*5*) Electrophoresis of AFLP fragments on an acrylamide gel. Different individuals have different restriction recognition sites and selective nucleotides, giving different AFLP fragments

be surveyed to find polymorphisms. AFLP also has an advantage in that it does not require probes or DNA sequence information. AFLP markers are inherited dominantly, and heterozygotes cannot be distinguished from homozygotes (as the allelic relationships of AFLP bands are difficult to determine). AFLP techniques have been used in various fields, most intensively in biodiversity analyses, trait-associated marker development, and linkage map development.

Single Nucleotide Polymorphism (SNP) Markers SNPs are literally single-base differences among individuals and represent 80 % of DNA polymorphisms. On average, one SNP exists every 1 kb in the human genome and every 170 bp between the

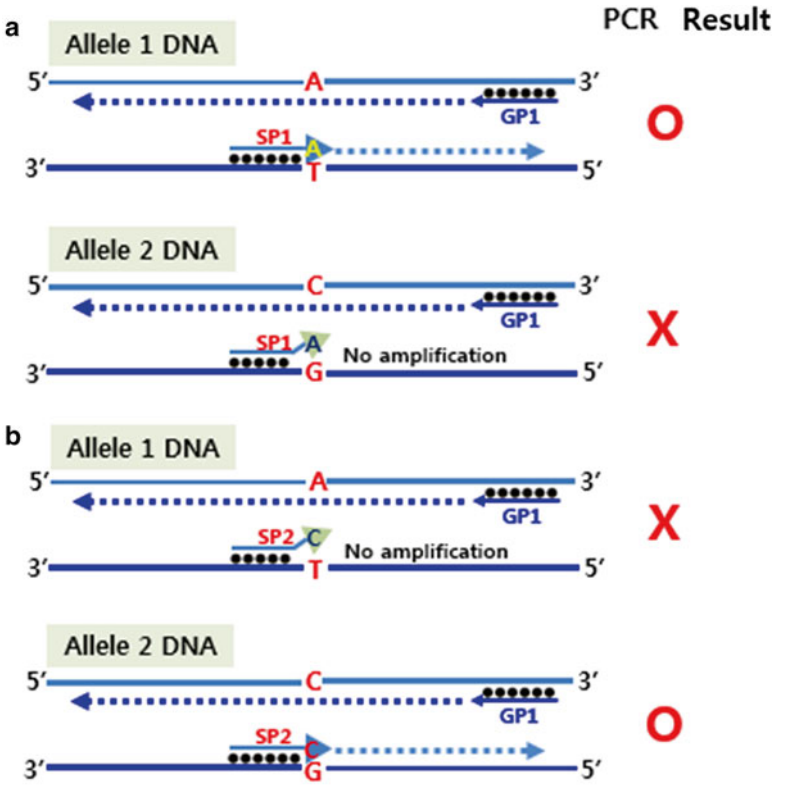

Fig. 2.6 Allele-specific PCR to analyze single nucleotide polymorphisms (SNPs). Using primer SP1, designed to have the allele "1" SNP at its 3′ end, PCR is performed. Because SP1 was based on its sequence, allele 1 amplifies well, while other alleles (e.g., allele 2) do not (**a**). However, if PCR is performed with primer SP2 with the allele 2 SNP sequence at its 3′ end, allele 2 will be amplified while allele 1 will not (**b**)

japonica and *indica* rice varieties. The most direct way of identifying SNPs in genomes is by directly comparing DNA sequences. Today, thanks to the accumulation of DNA sequence data, SNP markers can be developed *in silico* by analyzing existing databases. Once SNPs are identified, differences among individuals can be analyzed as follows.

1. Allele-specific PCR: When one of the two primers used in PCR has a 3′ end that does not complement the template DNA, DNA amplification is slower than usual. Therefore, if the SNP is set as the 3′-end of the primer sequence, one allele will amplify well, while different alleles will not, allowing easy allele distinction (Fig. 2.6).
2. 5′ Nuclease Assay: This method involves probes, such as TaqMan™, with a fluorescent label attached to the 5′ end, while the 3′ end has a chemical that inhibits

fluorescence (Fig. 2.7). The TaqMan™ probe is designed to be complementary to one allele. The probe does not cover the entire allele; instead it complements a small part that contains the SNP. If PCR is performed with the TaqMan™ probe attached, the 5′ exonuclease function of the PCR enzyme cuts the 5′ end, releasing the fluorescent compound and allowing it to light up. If PCR is attempted with different alleles, the 5′ end does not attach properly and is not cut by the enzyme, so no fluorescence is emitted. Therefore, the presence or absence of fluorescence reflects which allele the sample DNA has with respect to a given SNP. This method is expensive but can be automated.

3. DNA Chips: This method uses a DNA chip with various DNA sequences (about 25 bp-long) attached regularly on a glass or metal plate (Fig. 2.8). The DNA density of the chip is very high, and more than a million DNA fragments can be

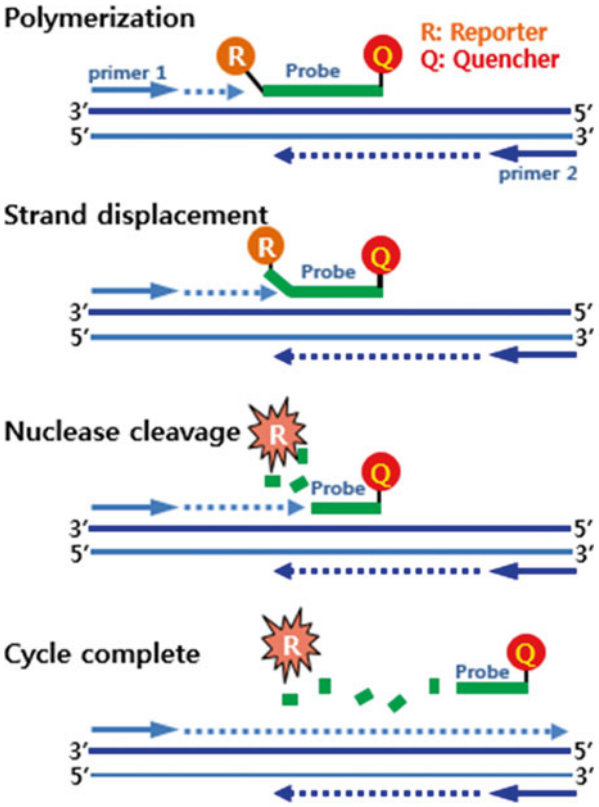

Fig. 2.7 TaqMan™ probe assay and 5′ exonuclease activity of the Taq polymerase used to analyze single nucleotide polymorphisms (SNPs). On the 5′ end of the probe, a fluorescent chemical (R) is attached. On the 3′ end, a substance that inhibits fluorescence (Q) is attached. When PCR is performed, the exonuclease activity of the DNA polymerase cleaves the 5′ fluorescent substance off, allowing fluorescence to occur. SNPs can be determined by light emission

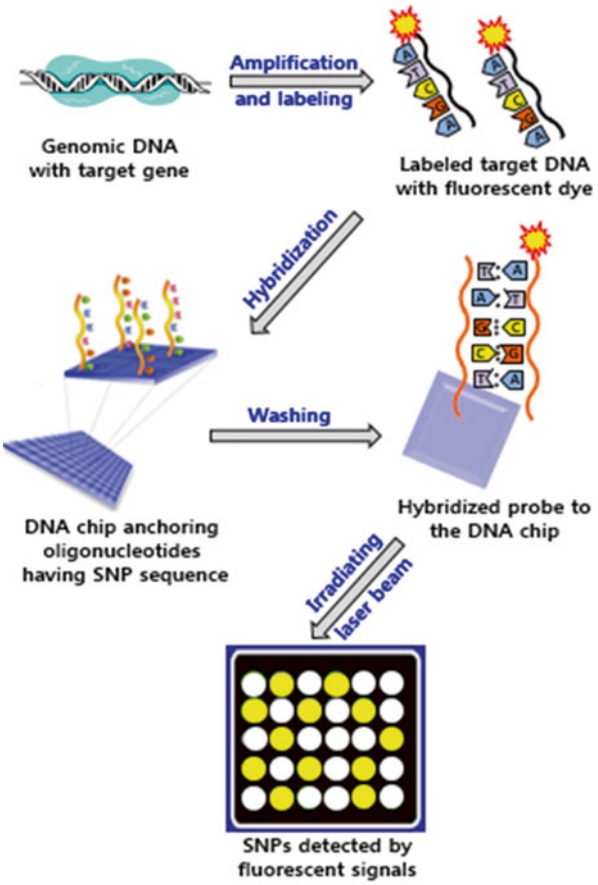

Fig. 2.8 DNA chips to analyze single nucleotide polymorphisms (SNPs). A DNA chip is prepared by attaching DNA fragments of ~25 bp in length from a species to a glass or metal plate. These fragments contain the SNP variants to be analyzed. Fluorescence-labeled probes are hybridized with the DNA fragments on the plate, and the DNA chip is washed. Finally, the probes are exposed to light of a specific wavelength to induce fluorescence

arranged per square centimeter. For the analysis, sections of the sample DNA containing SNPs must be amplified using PCR. Fluorescent dyes are attached to these fragments, which are used as probes to hybridize with the pre-prepared DNA chip. Then, the chip is washed. The probes that complement the chip sequences will attach firmly, while other allele fragments will be washed away. By analyzing the fluorescent spots, the alleles in the sample DNA can be determined. DNA chips are useful when analyzing species for which genome information is available.

Pros and Cons of Various DNA Marker Methods The ideal DNA marker would have the following characteristics: even distribution throughout the genome, high experimental reproducibility, fast and cheap analysis, little DNA requirement, and

Table 2.2 Comparison of DNA markers

Marker	Reproducibility	Technical difficulty	Inheritance	Sequence information
RFLP	High	High	Codominant	Not required
SSR	High	Medium	Codominant	Required
RAPD	Low	Low	Dominant	Not required
AFLP	High	Medium	Codominant	Not required
SNP	High	High	Codominant	Required

linkage with various phenotypes. Currently, no DNA markers meet all of these conditions; each technique has pros and cons (Table 2.2) that must be considered when planning experiments so the selected markers match the purpose of the experiment.

2.1.2 Tools for Trait-Linked Molecular Marker Development

2.1.2.1 Plant Material

A simple way to develop molecular markers linked to useful traits in agriculture is to find clear differences between allelic sequences of the target trait. However, because anonymous sequence differences between two individuals are not likely to be located in the target region, carefully prepared plant materials must be used to identify trait-linked molecular markers. In other words, because DNA sequence diversity often does not represent phenotype differences, plant materials must be prepared to develop trait-linked molecules, as described below.

Segregating Population This method is one of the easiest ways to prepare plant materials for developing trait-linked markers. Two plants with and without the target trait are crossed to develop an F_2 or backcross BC_1 population. The former is derived by self-pollination of F_1 individuals, whereas the latter results from crossing an F_1 hybrid with one of its parents. These populations contain individual plants segregating the target trait. Molecular markers linked to the trait can be developed by analyzing individuals that have the target trait and those do not. In other words, molecular markers linked to the target trait can be developed by first finding marker polymorphisms between two parents then analyzing the entire segregating population to find markers correlated to the phenotype.

Near Isogenic Lines (NILs) NILs are isogenic lines that have the same genomic background except at one locus that has different alleles. NILs are developed by repeatedly backcrossing a donor parent having a desired allele and a recipient (recurrent) line lacking the target allele (Fig. 2.9). By selecting individual plants having the desired allele in each generation, a line with both the recipient genome and the target trait is developed. Developing NILs is time-consuming and tedious. However, they are very useful for developing trait-linked molecular markers. For

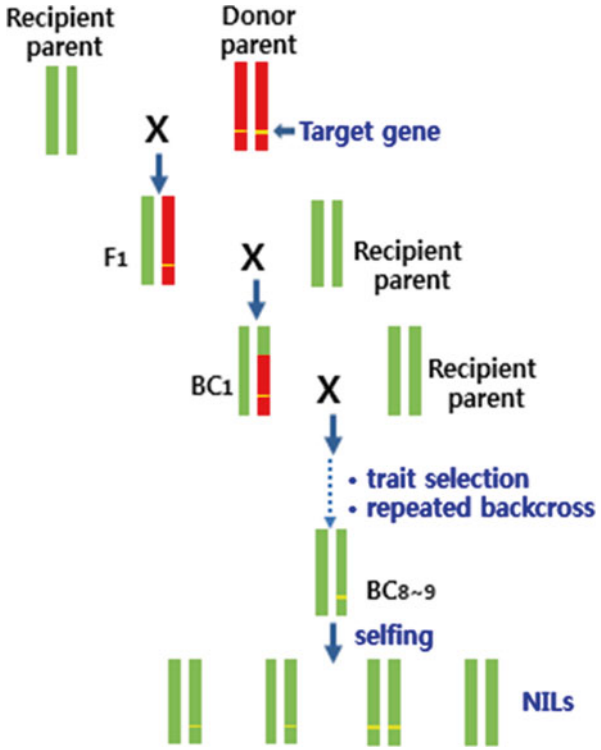

Fig. 2.9 Process of developing near isogenic lines (NILs). The F_1 is created by crossing a donor and a recipient parent. Recipient parents are repeatedly crossed with offspring that have the target allele to create NILs with the target allele in the recipient genome

example, if two individuals of a NIL set show DNA sequence difference, a DNA polymorphism between them is probably closely linked to the target trait, because their only genetic difference is near the target locus. NILs are often used for fine mapping of a specific locus and for physical mapping in map-based cloning of useful traits.

Recombinant Inbred Lines (RILs) RILs are obtained from F_2 individuals by successive self-pollination via single-seed descent (Fig. 2.10). First, the size of the F_2 generation should be determined based on the experimental goal. A single seed is chosen from a plant in each generation and propagated by self-pollination. After eight or more consecutive generations, individual plants from each family member derived from a given F_2 plant become genetically uniform. RILs are time-consuming to develop but are used frequently because their loci are genetically fixed and very useful for studying quantitative traits. Because RILs are genetically homozygous, experiments can be repeated in various labs around the world.

Double Haploid (DH) Lines DH lines are similar to RILs in that individual plants from each family derived from a given population are genetically uniform. DH lines

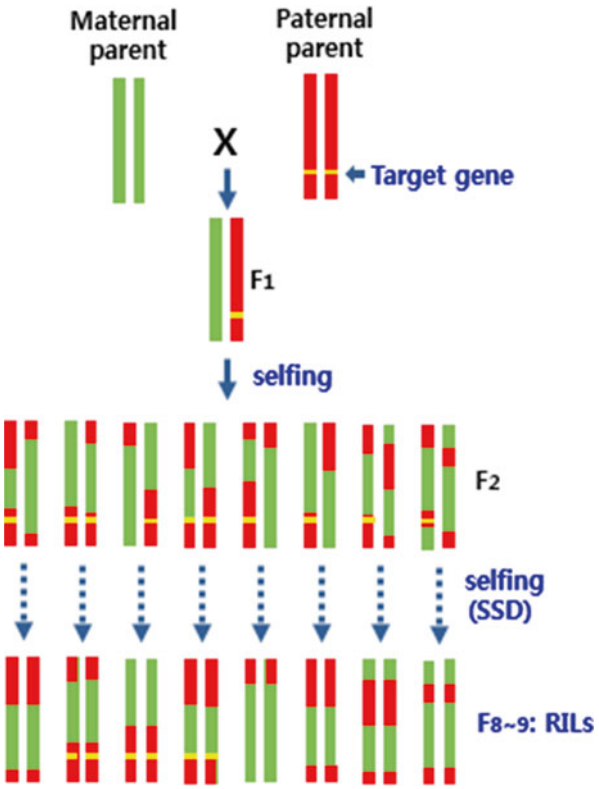

Fig. 2.10 Process of developing recombinant inbred lines (RILs). Two parents with different genotypes for the target trait are crossed to form F_1 plants, which are self-pollinated to get F_2. Single seed descent is used to obtain genetically fixed lines

are developed by culturing haploid tissue, such as the anther, microspore, or ovary. These cells become diploid either naturally or via colchicine application. DH lines have the same characteristics as RILs.

2.1.2.2 Phenotype Testing and Genetic Analysis

Agriculturally important characteristics include plant architecture, photosynthesis ability, seed size, flower-related traits, fruit color, taste, the presence of functional compounds, and abiotic resistance. Among these traits, some traits of disease resistance or color are controlled by only a few genes, and the phenotypes of these qualitative traits are relatively easy to determine, allowing accurate identification of the inheritance mode in these traits. However, traits like productivity and abiotic-stress resistance are controlled by multiple genes and show continuous phenotype variation, making the effects of the underlying genes hard to analyze and their inheritance mode are even more difficult to determine. Developing molecular markers for

such quantitative traits requires an accurate phenotyping method. To investigate the inheritance of a quantitative trait, two different inbred lines are crossed to make the F_1, which is then self-pollinated or backcrossed with one of the parents to develop the F_2 or BC_1. In each population, the phenotype is analyzed to estimate the number of genes controlling the trait. Even so, the inheritance mode is hard to identify when more than three genes are involved in the trait. In these cases, quantitative trait locus (QTL) mapping can help identify the number of involved genes and their roles.

2.1.2.3 Linkage Mapping and Genome Information

Linkage Mapping (Genetic Mapping) When two parents with distinct alleles at many loci are crossed, the allelic combinations on individual chromosomes can change in subsequent generations because of cross-over events during meiosis. These novel combinations yield descendants with unique phenotypes different from their parents. However, when two loci are located closely on the same chromosome, the probability of cross-over events between them falls, and the recombinant genotype becomes relatively rare. The cross-over rate increase in proportion to the distance between genes, so cross-over rate data allow the estimation of loci distances on the chromosome. A genetic map constructed by this way shows the relative locations of morphological or molecular markers. On such maps, one map unit is defined as having a cross-over rate of 1 % and is called a centimorgan (cM). If a genetic map is available, the genotype/phenotype correspondance of individuals in a segregating population can be calculated by comparing the phenotype and the marker genotype.

Genome Information A genome is defined as all the genetic information of a species; it includes both nuclear and cytoplasmic genetic information. The unique genetic information of an organism affects its phenotype. Thanks to rapid technological developments, genome analysis has become a routine way to study gene functions and evolution of animals, plants, and microorganisms. Genome analysis can be used both within a species and among related species. Rather than focusing on the function and expression patterns of each gene, genome research surveys the entire genome sequence to analyze gene structure and find a large scale regulatory mechanisms of the genes.

2.1.3 Glossary

Allele-Specific PCR One way to test SNPs. If PCR is performed with a primer whose 3′ end complements the SNP, an allele that perfectly matches the primer will be amplified while others will not. This principle is used to distinguish SNP alleles.

Allele One of two or more variants of a gene can reside at the same locus on a chromosome. If the loci on homologous chromosomes have the same allele, the

individual is said to be homozygous at that locus. If the alleles are different, the individual is heterozygous for the trait.

Amplified Fragment Length Polymorphism (AFLP) A DNA marker that combines the advantages of both RFLP and RAPD. Sample DNA is digested with two restriction enzymes, and adaptors are attached to the fragment ends. Primers specific to the adaptors are used to amplify DNA using PCR. The PCR products are separated by electrophoresis to observe individual polymorphisms.

Centimorgan (cM) The genetic distance along a chromosome with a 1 % probability of cross-over between two molecular markers or loci.

Codominant Marker Markers that can distinguish the homozygote and heterozygote states.

Codominant In heterozygotes, when the contributions of both alleles are seen in the phenotype.

Crossing Over The exchange of paternal and maternal segments of homologous chromosomes during meiosis. It causes progeny to have different allele combinations from their parents.

DNA Chip One way to test SNPs. Artificial 25 bp oligonucleotides are attached to glass or metal plates. Fluorescently-labeled probes of DNA fragments are hybridized to determine which alleles have a given SNP.

DNA Hybridization A process that combines two complementary single-stranded DNA molecules to form double-stranded molecule through base pairing.

Dominant Marker A marker that cannot distinguish the homozygous and heterozygous states.

Dominant A genetic phenomenon in which the phenotypes of dominant homozygotes (AA) and heterozygotes (Aa) are the same.

Doubled Haploid Lines (DH Lines) A plant that is created by culturing haploid cells (anther, microspore, or ovary cells) and converted diploids naturally or by adding colchicine. DH lines have the same genetic characteristics as RILs.

Expressed Sequence Tags (ESTs) 5′ and 3′ end sequences of clones in cDNA libraries. ESTs are 300–500 bp single-strand mRNA sequence reads derived from genes expressed in a given tissue and/or at a given developmental stage. The technique is used in functional genomics.

5′ Nuclease Assay A way to test SNPs that uses probes, such as TaqMan™. A fluorescent chemical is attached to the probe's 5′ end, while on the 3′ end bears a fluorescence-inhibiting chemical. During PCR, the 5′ exonuclease function of the PCR polymerase cuts the inhibiting 3′ end, allowing fluorescence to be generated to indicate SNPs.

Functional Genomics A set of methods to identify gene function using genome sequence data acquired from structural genomics research. It includes methods like high-throughput sequencing of expressed genes (ESTs), gene expression analysis using DNA chips, and two-dimensional electrophoresis to differentiate proteins.

Genetic Marker A marker that is closely linked with useful traits, such as disease resistance or abiotic stress resistance.

Genetic or Linkage Map A map that shows the relative locations of genetic or molecular markers calculated using cross-over rates. Also called gene linkage-group map.

Genome The complete set of genetic information found in an species (includes both nuclear and cytoplasmic genetic information).

Isozyme A type of protein marker. Isozymes of an enzyme have the same activities but can be distinguished by different electrophoresis speeds.

Library A gene set that contains clones of certain DNA fragments from a target crop. Various types exist, such as a genomic library of the entire genomic DNA of a sample, and a cDNA library of information about expressed genes.

Linkage The genetic proximity of two or more genes on a chromosome.

Locus The location of a gene on a chromosome or chromosome map.

Microsatellite A 1–6 bp DNA repeat that is usually repeated fewer than 100 times. Also called simple sequence repeats (SSRs).

Minisatellite A 9–100 bp random sequence that is usually repeated fewer than 1000 times.

Molecular Marker A protein or DNA marker used to easily distinguish a target trait.

Morphological Marker A marker that can distinguish the genotype by the form of the phenotype, instead of requiring biochemical or molecular biology techniques.

Near Isogenic Lines (NILs) Lines that have the same genotype except at one locus that has different alleles.

Physical Map A map that showing the positions of sequence features, including genes. Physical maps are generated using DNA sequence data.

Polymerase Chain Reaction (PCR) A technique for amplifying a specific DNA sequence. A cycle of DNA denaturation → primer annealing → polymerization (using a heat-resistant DNA polymerase) is repeated to amplify DNA within a machine.

Polymorphism The phenomenon in which more than one different allele exists at the same locus. Polymorphism can be observed at the phenotype, protein, or DNA level.

Primer A short nucleotide sequence that acts as the starting point of DNA replication. It complements the template DNA strand.

Probe A DNA fragment that complementarily bonds to a specific DNA sequence. Probes are labeled with isotopes or fluorescent dyes to easily identify homologous genes on DNA (or RNA) blots or DNA chips.

Random Amplified Polymorphic DNA (RAPD) A type of DNA marker that can be developed without prior sequence information about the target gene. A random sequence primer (usually 10 bp long) is used to perform PCR; electrophoresis of the amplicons reveals polymorphisms among individuals.

Recombinant Inbred Lines (RILs) RILs are obtained by successive self-pollination from F_2 individuals via single-seed descent. RILs are often used to find quantitative trait loci (QTL mapping).

Restriction Endonuclease An enzyme that recognizes a specific base-pair sequence (usually 4–8 bp in length) and cuts the DNA double strand.

Restriction Fragment Length Polymorphism (RFLP) DNA polymorphism revealed by digesting DNA with restriction enzymes, followed by electrophoresis. The term also denotes technology that confirms such differences using Southern blotting and DNA probes.

Sequence Tagged Site (STS) A type of molecular marker that uses the DNA sequence near a target gene as a primer for PCR to find differences among amplified regions.

Simple Sequence Repeat (SSR) (See microsatellite.)

Single Nucleotide Polymorphisms (SNPs) A single nucleotide difference between individuals of a species.

Southern Blot A method of transferring DNA fragments that were generated by restriction enzymes and electrophoresed onto a membrane; often used to test for the presence of homologous genes in a genome.

Structural Genomics A research method to understand genetic phenomena from a macroscopic perspective by studying the genome's molecular structure. Methods like high-density gene mapping and genome sequencing are included in structural genomics.

Variable Number of Tandem Repeats (VNTRs) Polymorphisms based on differences in minisatellite repeats. VNTRs can be confirmed using DNA probe hybridization or PCR.

2.2 Developing Trait-Linked Molecular Markers

When a target trait is controlled by a single gene, molecular markers linked to the trait can be developed by generating a linkage map. To check whether a molecular marker and target trait are linked and to calculate their genetic distance,

co-segregation analysis must be performed. A mapping population such as F_2 or BC_1F_1 should be constructed, and then the phenotype and molecular marker genotype are analyzed. The number of recombinant individuals is counted, and the genetic distance between the molecular marker and the target gene can be calculated in cM units to generate a genetic linkage map.

However, because only a small fraction of the genome contains a given trait, making a genetic linkage map of an entire plant genome could be a waste of time and energy. Additionally, saturating a genetic linkage map is not guaranteed to lead to the development of a molecular marker closely linked to the target gene. Therefore, methods for developing molecular markers that are more efficient have been devised. Prior to the advent of sufficient biological information like genome sequences, researchers found efficient ways to develop molecular markers without a large amount of data.

2.2.1 When Genetic Linkage Maps or Biological Data Are Not Available

When biological data were insufficient, NILs developed during plant breeding were often used. If such plant material could not be prepared, bulked segregant analysis (BSA) techniques were used to develop molecular markers. Molecular markers such as RAPD or AFLP can be used.

2.2.1.1 Near Isogenic Lines

Theoretically, the genomes of NILs differ only in the region near the target trait. NILs can be developed using backcrossing (Fig. 2.11).

1. A ‘donor’ (P1) with the target gene is crossed with the ‘recurrent’ parent (P2) to obtain the F_1 generation.
2. The F_1 and P2 are backcrossed to make the next generation. This backcrossing process is usually repeated 4–5 times.
3. In each generation, individuals with the target trait are chosen to be backcrossed again. Thus, all genetic ‘background’ except the target gene is replaced with the P2’s genetic information.
4. A line that is homozygous for the target gene is selected to develop an NIL through 1–2 rounds of self-fertilization.
5. To obtain NIL, plants should be advanced to the BC_4F_2–BC_5F_3 generation. NILs can be developed with fewer backcrosses when RILs or introgression lines (ILs) already exist.

Developing molecular markers linked to a target gene using NILs begins with comparing the genotype of the P1, P2, and NIL using RAPD or AFLP markers. There are three possible outcomes: (1) the P1, P2, and NIL all have the same genotype; (2) the P2 and NIL have the same genotype but not the P1; or (3) the P1 and

Tip
Making NILs requires a systematic way to select individuals with the target gene for each backcross generation. If the target gene is a male-sterility restoring or fruit-color gene, individuals can be easily selected by observing the anther or fruit, respectively. However, if the target gene is related to disease resistance, a disease-screening experiment must be performed to identify individuals that carry the gene. If the target gene makes a functional compound, the amount of product can be measured experimentally. Whether the target trait is dominant or recessive is also an important factor. If the P1's genotype is AA and the P2's genotype is aa, the genotype of the BC_1F_1 generation segregates to Aa:aa = 1:1. If the target gene is dominant, the heterozygote will show the target trait, so individuals with the target trait can be selected directly to advance to the next generation. However, if the target gene is recessive, none of the individuals will show the target trait in BC_1F_1, so progeny must be tested to find the heterozygotes.

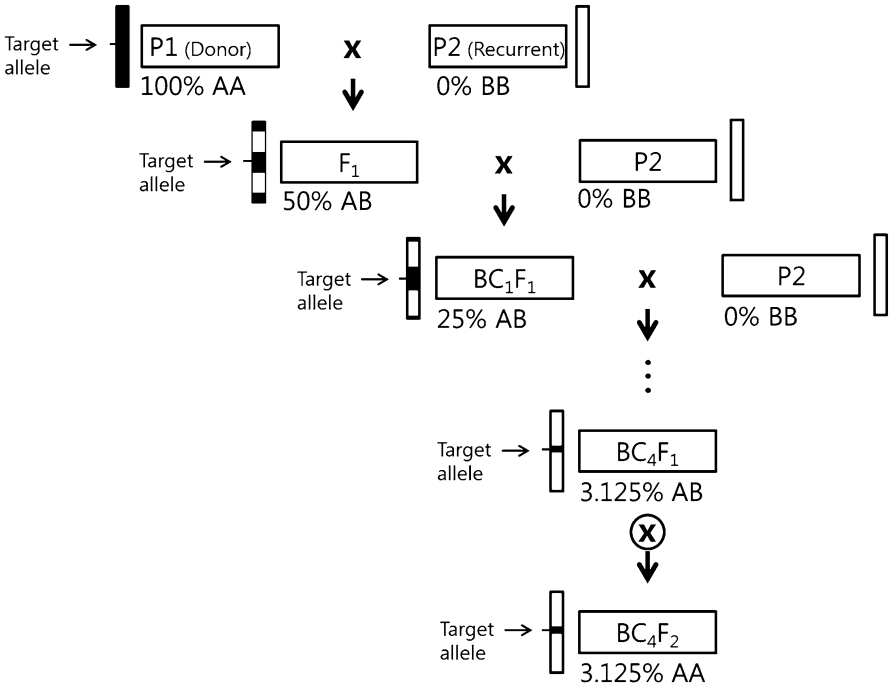

Fig. 2.11 Near isogenic line (NIL) development scheme. For each generation, the *bar* indicates a chromosome; *black* represents a genetic segment from the donor and white a segment from the recurrent. The amount of genetic data inherited from the donor at each generation is shown as a percentage at the *bottom* of each generation. The genotype of the target gene is shown as AA (homozygote) or Aa (heterozygote)

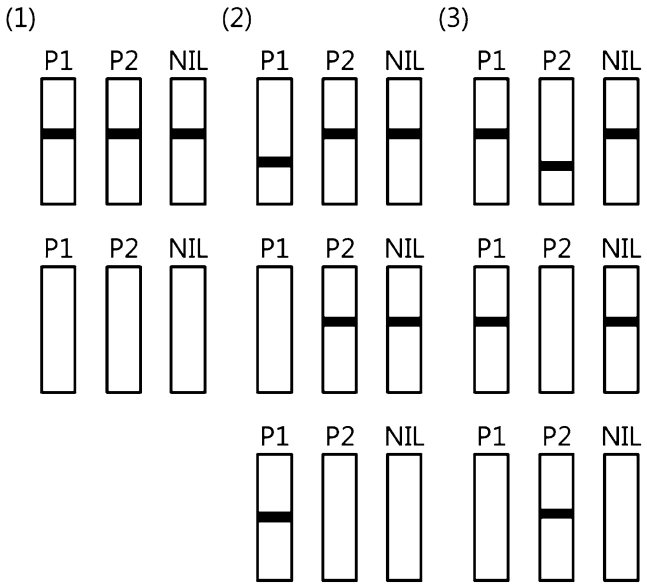

Fig. 2.12 Example of analysis results for the donor (P1, genotype AA), recurrent (P2, genotype aa), and near isogenic line (NIL, genotype AA). The picture shows each possible PCR result for molecular markers. The *boxes* represent lanes in an agarose/acrylamide gel after electrophoresis, with bands shown as *black lines*

NIL have the same genotype but not the P2 (Fig. 2.12). In the first case, because the three lines have no polymorphism, it cannot be used in co-segregation analysis. In the second case there is polymorphism, but because the genotypes of the P2, which lacks the target gene, and NIL, which has it, are the same, the marker is probably far from the target locus or not linked to it. In the third case, there is polymorphism and, because the marker genotype of the NIL is the same as that of P1, the marker is probably linked to the target locus. After choosing markers that correspond to the third category, co-segregation analysis can be performed to check the degree of linkage and, ultimately, the selected markers can be used in experiments.

NILs are very useful in genetic research, and their utility is decreased by their long development times. Additionally, in the case of a BC_4F_1 generation, 3.125 % of its genome should come from the P1 plant, but that genomic fraction can exist across several locations that are not linked to the target gene. Because, it is dispersed throughout the genome and may cause false positive errors to occur (Fig. 2.13).

2.2.1.2 Bulked Segregant Analysis (BSA)

BSA is a method that can be used when experimental material like NILs is not available. To find molecular markers linked to a target gene, genomic DNA of several plants with the same phenotype is bulked into one pool. Two bulked pools of

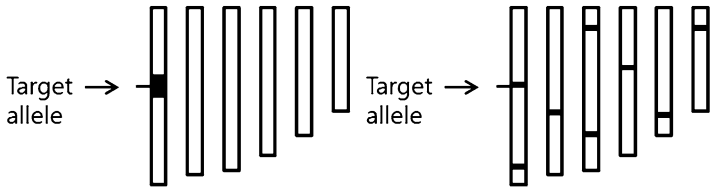

Fig. 2.13 Explanation of false positive errors in near inbred lines (NILs). The *vertical bars* represent chromosomes; the *black areas* originate from the donor. At *left*, the donor-origin areas are concentrated near the target gene, while at *right*, the donor-origin genetic material is dispersed to places unlinked with the target gene

segregants differing for one trait will differ only at the locus harboring that trait. BSA can be performed via the following steps (Fig. 2.14).

1. Cross a target gene-containing P1 line and gene-lacking P2 line to generate the F_1 generation, and self-pollinate F_1 to get the F_2 generation.
2. Determine the phenotype for the target gene in each individual and extract DNA. Dilute the DNA to the same concentration, then mix the DNA samples of the same phenotype to make a DNA 'bulk' for each phenotype.

Tip
When performing BSA, two things must be considered: the minimum individual number to comprise a DNA bulk and how to distinguish homozygotes from heterozygotes with the same phenotype. The minimum individual number for a bulk is closely related to the maximum probability that a molecular marker will actually be linked to the target locus. If the individual number in a bulk is *n*, the probability that a molecular marker will be linked with the target gene is calculated as $2\times\left(\frac{1}{4}\right)^n\times\left\{1-\left(\frac{1}{4}\right)^n\right\}$. When the individual number in a bulk is 4, the probability is 1 out of 100; for n=10 it is 2 out of a million, and for n=15 it is 2 out of a billion. DNA bulks usually consist of DNA from 10 to 15 individuals. Choosing homozygotes requires test crossing in later generations.

If one assumes that both parents have a homozygous genotype at all loci, the genotypes of all F_1 individuals will be the same. Consider two other loci, B and C, that are far from the target gene. After performing BSA and making DNA bulks, there must be an AA bulk and an aa bulk. While the allele frequency for gene A in the AA bulk is A:a=1:0, for other loci that are not linked to A, the genotype frequency will be BB:Bb:bb=CC:Cc:cc=1:2:1. Thus, the allele frequency for other loci is 1:1, showing heterozygotes as a result (Fig. 2.15).

Like NILs, when DNA bulks are used in RAPD or AFLP marker analysis, there are three possible results: (1) the AA bulk and aa bulk are both heterozygotes, (2) the AA bulk and aa bulk are both homozygotes but do not have polymorphism, and

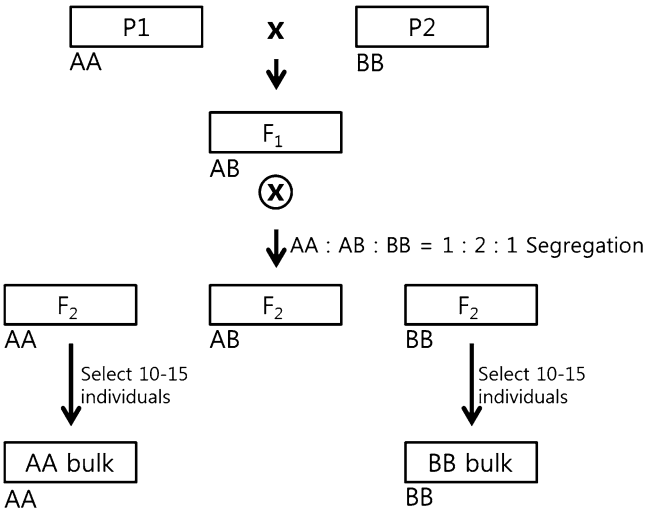

Fig. 2.14 Bulked segregant analysis (BSA) scheme, showing important steps in the progress. The genotype of each generation is written as AA (homozygous for target allele), Aa (heterozygous), and aa (homozygous for non-target allele)

(3) the AA bulk and aa bulk are both homozygotes and show polymorphism. As with NIL, molecular markers with the third outcome are selected, and co-segregation analysis is performed to develop trait-linked molecular markers (Fig. 2.16).

Because BSA can be performed with only the F_2 generation, it is one of the most efficient ways to develop molecular markers. BSA is used in molecular marker technologies like RAPD and AFLP, and applications are expanding to DNA chips or next generation sequencing (NGS) technology.

2.2.2 When Genetic Linkage Maps or Biological Data Are Available

Many genes have been identified and their functions were determined in various organisms. This biological information can be used to discover the functions of genes in other species. Thanks to the development of molecular genetics, the genetic linkage maps for many species have been saturated, and physical maps based on bacterial artificial chromosome (BAC) libraries have contributed greatly to elucidating the functions of previously unknown genes and developing new molecular markers. Now, NGS technologies are allowing entire genomes to be sequenced as references, and resequencing individuals of the same species has become common.

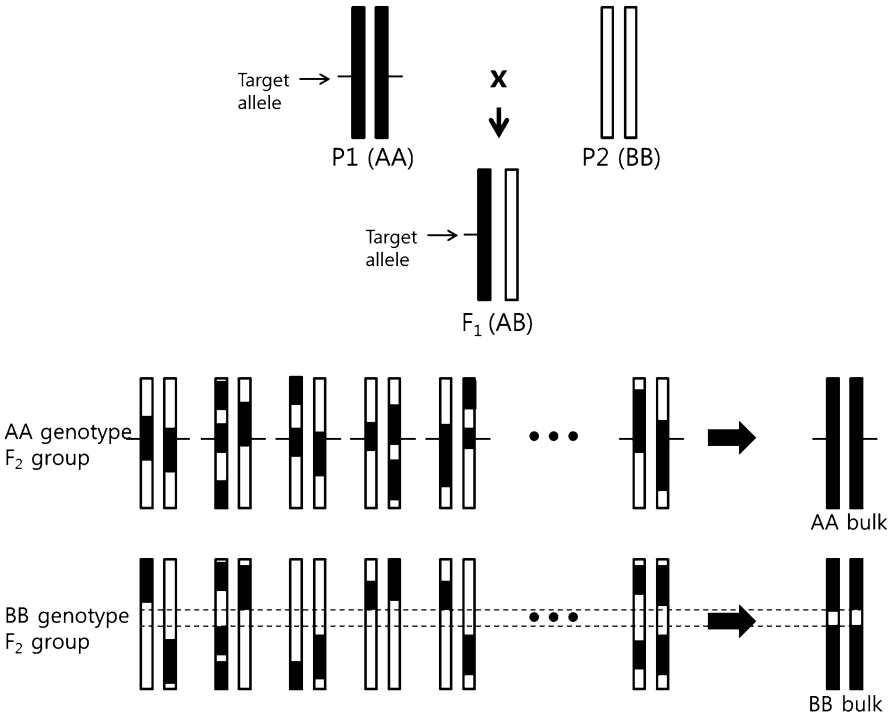

Fig. 2.15 Diagram of P1, P2, F1, and F_2 diploid group in bulked segregant analysis (BSA). The *vertical boxes* represent the chromosome on which the target locus is located; the *black area* indicates genetic material from P1 and the white area from P2. The lower picture is an example of possible genetic backgrounds for genotypes AA and aa. When these individuals are picked to make a DNA bulk, all areas of the chromosome except the target locus become similar

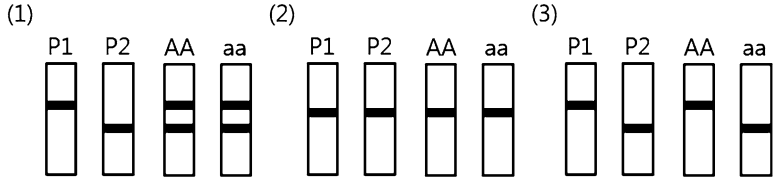

Fig. 2.16 Example of molecular genotype analysis performed on P1, P2, the AA bulk, and the aa bulk in bulked segregant analysis (BSA). The picture depicts possible genotype analysis results for PCR-based molecular markers. The *vertical boxes* represent post-electrophoresis agarose/acrylamide gel lanes, and the *black lines* signify amplified bands

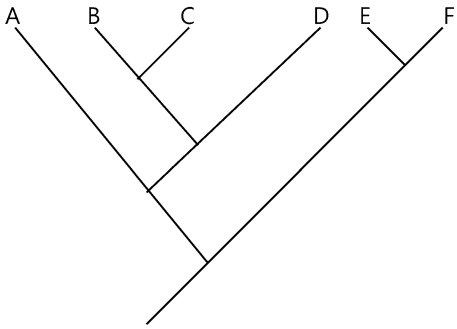

Fig. 2.17 Phylogenetic tree of six species originating from a common ancestor. Because speciation occurred early between the *A–D* and *E–F* groups, they have little in common. In contrast, species *B* and *C* are similar to one another, as are species *E* and *F*

Such abundant biological information allows new molecular markers to be developed with speed and accuracy that was only in the realm of imagination before.

Methods to develop molecular markers depend on whether existing biological information can be used or not. When there is abundant biological information about a crop plant, the data can be used to develop molecular markers via a candidate gene or comparative genetics approaches.

2.2.2.1 Using Comparative Genetics

Comparative genetics is based on evolution as a theoretical background. An important assumption of evolutionary theory is that any group of organisms will have a common ancestor (Fig. 2.17). Species that recently shared a common ancestor have similar numbers of chromosomes, and the gene loci are also similar (a feature termed synteny). The more distantly related two species are, the less commonality in their DNA, so comparative genetics is usually applied to species in the same genus or, in some cases, more broadly to species within a family.

Tip
Synteny describes the phenomenon in which two phylogenetically close species share common genes and chromosomes from their ancestor and, therefore, are similar. For example, if a gene for fruit color exists between molecular markers 1 and 2 in species B, the same gene will probably also appear between markers 1 and 2 in a closely related species C because of synteny (Fig. 2.18).

Early comparative genetics began by comparing the genetic maps of different species. This process requires the availability of a common marker that can be used in at least two genetic maps. The comparison becomes more accurate as more markers are shared by the individual genetic maps (Fig. 2.19). If we consider two linkage groups

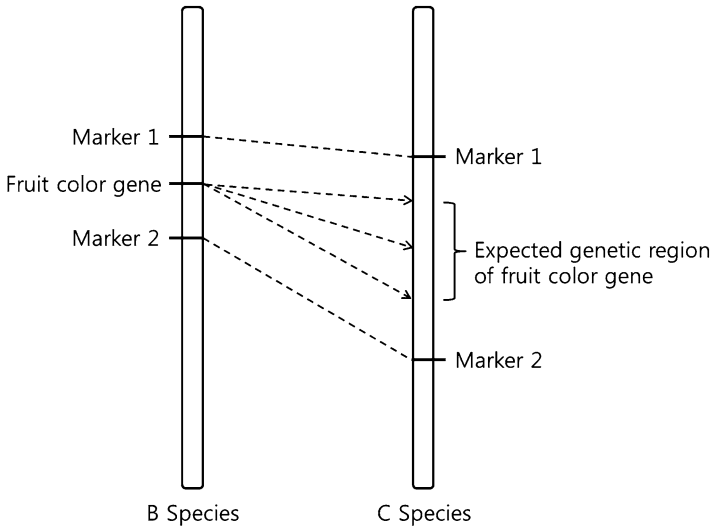

Fig. 2.18 Chromosome map used in comparative genetics showing synteny between two phylogenetically close species

with only one common molecular marker, that marker can identify the syntenic genetic region but cannot show the orientation of markers (Fig. 2.19a), which requires at least two common markers to determine (Fig. 2.19b, c). The degree of synteny can differ with genetic distance. The closer two species are phylogenetically, the greater the synteny in both local and large scales, for example at the chromosome level (Fig. 2.19d). However, if two species are distantly related, synteny is conserved only in small regions, but not in large scale ('macro synteny' is not conserved) (Fig. 2.19e). Therefore, if only micro synteny is observed, more common markers may be needed to find the genetic region containing the target gene. Also, analyses must be performed cautiously even in syntenic areas, which may contain occasional rearrangements.

If complete genome or EST sequence data exist for the species that is being used for comparison, different methods of research become possible. If the entire genome sequence is known for a model species and only EST sequence data exist for the target plant, the ESTs can be used as common markers because they originate from expressed genes. Therefore, EST sequences are highly likely to be conserved between the two species, and similar EST sequences can be assumed to be present in the genome of a model plant. When comparing species that are distantly related, the tblastx algorithm should be used, while the blastn algorithm is preferable when investigating closely related species. The tblastx algorithm finds similar amino acid sequences; therefore, it may not work accurately when the genome contains many paralogs. Because closely related species have more similar orthologs than paralogs, the blastn algorithm, which compares nucleotide sequence similarity, is better for them.

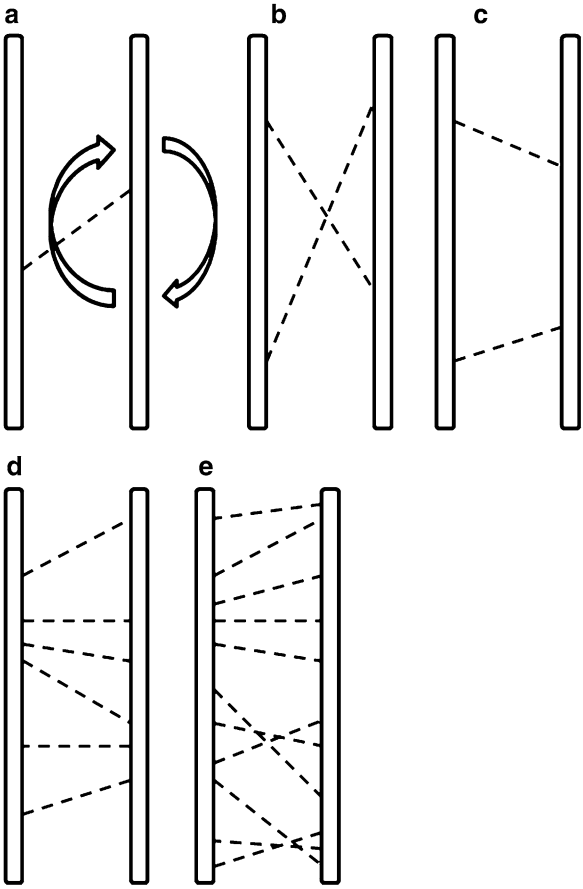

Fig. 2.19 Example of a comparative genetics analysis of two different species. The picture shows a comparative linkage map for cases when (**a**) there is one common marker, (**b**, **c**) there are two common markers, and (**d**, **e**) there are many common markers. When more than two common markers exist, two linkage groups can be compared. For example, in (**b**), the two linkage maps are reversed. In (**d**), both macrosynteny and microsynteny can be observed. In (**e**), the species evolved separately and experienced rearrangement of the chromosomal genes, leaving only microsynteny

Tip

The Basic Logical Alignment Search Tool (BLAST) is one fundamental program for analyzing nucleotide sequences. BLAST compares a submitted sequence (query) with others in a database (DB) and aligns the query with any similar sequences found. BLAST has many embedded algorithms, including blastn, blastp, blastx, tblastn, and tblastx. Their uses will vary by DB and query. Blastn is used when the query and DB are both nucleotide sequences, while blastp is used when both are amino acid sequences. The other three algorithms artificially translate nucleotide sequences to amino acid sequences to compare the query with the DB: blastx uses an expressed sequence query against an amino acid DB; tblastn uses an amino acid query against an expression sequence DB, and for tblastx, both the query and DB are amino acid sequences.

2.2.2.2 Finding Functional Candidate Genes

Candidate genes can be divided into functional and positional candidates. In plants, the sequence, structure, and function of various genes have been discovered through research using model plants such as *Arabidopsis*, rice, and tomato. Assuming that all species have a common ancestor (Fig. 2.17), a gene with a known function in one species will likely have a similar function in another. This inference justifies the development of molecular markers using functional candidate genes. The theoretical range for such comparisons can be within a kingdom, but candidate genes are usually found within lower-level taxa, such as a genus or family, because genetic similarities are more easily found.

A similar gene structure indicates that the expressed three-dimensional structures of two proteins are similar and that the amino acid sequences are identical or similar. Consider a gene that determines fruit color in species A. The following strategies can be used to find functional candidate genes in species B. If species B has its entire genome known or at least an EST database, similar genes in B can be found by querying the fruit color gene for A in the database of species B. For such a search converting the DB and query base-pair sequences to amino acid sequences, then comparing the sequence similarity instead of identity is convenient. As noted above, the tblastx algorithm is appropriate for this purpose.

Candidate genes found in EST or whole-genome DBs by BLAST should be confirmed experimentally. To find the entire sequence for the gene, primers should be designed to identify the 5′ and 3′ UTR regions. If the resulting base-pair sequence shows high similarity with the EST or whole-genome DB, the candidate gene is confirmed to be present in the genome.

Tip
If identity analysis is a process of confirming whether two sequences are the same, similarity analysis is comparing whether the two sequences are in the same ‘group’. Even if the species have their origin in one species, as they each go through their own evolutionary process, many mutations accumulate on the gene. Let us assume that if gene functions are impaired, there is a disadvantage in survival. If a mutation happens to change the amino acid sequence, usually the resulting enzyme cannot function properly; therefore amino acid-changing mutations are given negative selecting pressure. In contrast, ‘silent’ mutations that do not change the amino acid sequence can survive to spread their genes. In such cases, the identity of the corresponding nucleotide sequences can be very low, but the similarity of the amino acid sequence can be very high.

2.2.3 *General Strategies for Developing Molecular Markers*

Modern comparative genetics does not simply compare genetic maps or sequences. Rather, all possible information is used to develop molecular markers. The genomes of many species are being analyzed and massive genome information is accumulating thanks to the rapidly decreasing costs of NGS. Inexpensive systems that analyze gene expression by sequencing transcriptomes with NGS technology are already available. Enormous genetic data, with functions identified, can be found at NCBI, and the rate of development of analytical instruments for the field of bioinformatics is amazing.

To develop target-gene-related molecular markers, the phenotype information and molecular marker genotype should be known for each individual in a segregating population. Methods of analyzing the marker genotypes have been in great progress, and using NGS to analyze biological information is now a necessity. Commercial bioinformatics programs provide simple interfaces for researchers with little experience in this area, but filtering and analyzing the data are time consuming. In the big-data era, analyzing a large amount of data with quick homemade scripts will become an essential skill for developing molecular markers.

2.3 Case Studies on Developing Molecular Markers

2.3.1 *Using Bulked Segregant Analysis (BSA):* Phytophthora *resistance in pepper (Liu et al. 2014)*

2.3.1.1 Preliminary Research and Background

Resistance of pepper against *Phytophthora capsici* is controlled by QTLs. Numerous sources of resistance have been reported, for example in *Capsicum annuum* CM334, PI201232, AC2258, and Perennial. The inheritance patterns of resistance vary depending on disease screening conditions and the water mold isolate. However, several studies have reported that a major QTL is located on chromosome 5 of pepper (Bonnet et al. 2007; Quirin et al. 2005; Thabuis et al. 2004).

2.3.1.2 Plant Materials and Phenotype Analysis

Many different kinds of mapping population have been used for QTL analysis, such as F_2 or BC. However, families of RILs have been the main choice, because of their homozygosity and accumulated recombination. The effects of

environmental variation on quantitative traits can be much reduced by assessing several plants of the same genotype instead of a single plant. The YT population, consisting of 128 RILs at the F7–F9 generation of an intra-specific cross between *C. annuum* 'YCM334' and 'Tean' was used for inheritance analysis and marker development (Truong et al. 2012). *Capsicum annuum* 'YCM334' was the resistant parent and 'Tean' was susceptible. This study hypothesized that *Phytophthora* resistance in pepper acts as a monogenic trait with resistance being dominant over susceptibility under low disease-pressure conditions in which a low *Phytophthora* concentration (3×10^4 zoospore/mL) was used. A segregation ratio of 65 (resistant) and 78 (susceptible) plants was obtained by resistance screening of the YT RIL population.

2.3.1.3 Development of Molecular Markers

To develop molecular markers linked to the major QTL on chromosome 5, BSA-SFP (BSA-single feature polymorphism) was performed. For BSA, equal amounts of genomic DNA from 20 resistant and 20 susceptible lines selected from YT RILs were separately bulked and treated by DNase. The labeled DNA pools were hybridized on Affymetrix chips containing 30,000 ESTs, with five replicates for each DNA pool. Two candidate EST-based unigenes with different Dstat values (≥ 3 or ≤ -3) between resistant and susceptible were identified. To develop molecular markers using these two ESTs, BLAST searches against the '*Capsicum annuum* Database' (http://peppergenome.snu.ac.kr) were performed. Several scaffold sequences were obtained and tested between YT RIL parents. Only one primer, 002466, was polymorphic in YT RILs. Co-segregation analysis of this primer showed 11 recombinant lines out of 135 lines. This marker was designated PhytoSNP5. The linkage map of chromosome 5 was constructed using developed markers in the YT RIL mapping population. As the result, the LOD score between PhytoSNP5 and IBP557 was the highest.

2.3.2 *Using Candidate Genes: Markers from Tomato Linked to* Cmr1 *(Kang et al. 2010)*

2.3.2.1 Preliminary Research and Background

The *Cmr1* gene, which has a single dominant phenotype, confers resistance against *Cucumber mosaic virus* (CMV) in pepper. Previously, Kim et al. (2004) reported three CAPS markers (CAPS-A, CAPS-B, and CAPS-C) linked to *Cmr1* but not the gene's map location.

2.3.2.2 Plant Materials and Phenotype Analysis

The mapping population used in this study was a *C. annuum* 'Bukang' F_2 population derived from self-pollinated F_1 'Bukang' plants. CMV screening was performed in the F_2 population using the CMV-Kor strain. Two varieties of *C. annuum*, 'Bukang' and 'Jeju', were used as negative and positive controls, respectively. The susceptible plants showed typical CMV symptoms of vein clearing mosaic and leaf distortion. Resistant plants showed no symptoms. The segregation analysis of resistance and susceptibility in the F_2 'Bukang' population showed 236 resistant plants and 73 susceptible ones, fitting the 3:1 Mendelian segregation ratio (P=0.7326).

2.3.2.3 Development of Molecular Markers

CAPS markers linked with the *Cmr1* gene were determined to be about 3 cM away in the 'Bukang' F_2 population. To locate the *Cmr1* gene in a pepper linkage map, the CAPS-A and CAPS-B markers were mapped using the AC 99 F_2 population. CAPS-A is located in the centromeric region of LG2 near TG31A (Fig. 2.20). The *Tm-1* gene, a *Tomato mosaic virus* resistance (ToMV) gene, is located in the centromeric region of tomato chromosome 2, and *Cmr1* is in a syntenic region. Therefore, the *Tm-1* sequence was used as a candidate gene for *Cmr1*. One pepper

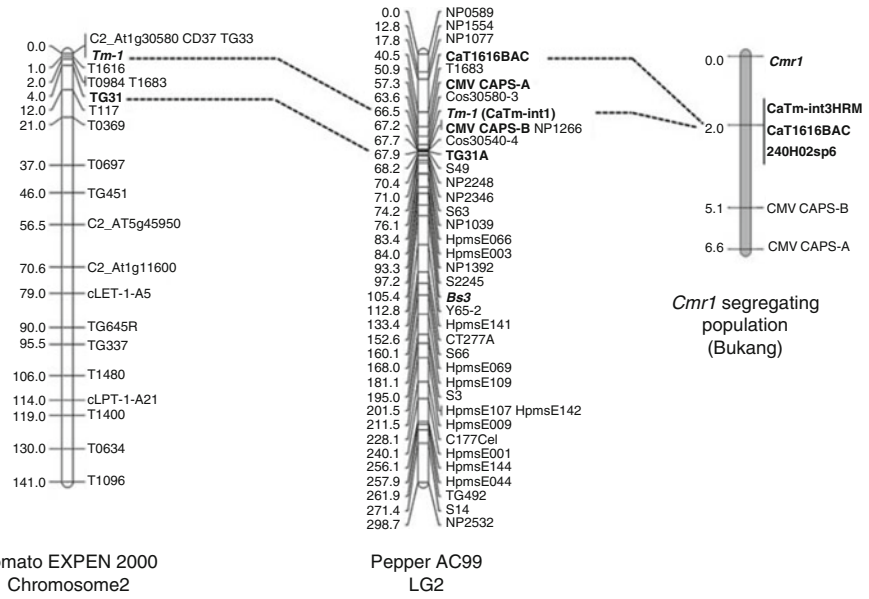

Fig. 2.20 Comparative analysis between tomato and pepper maps. The *Cmr1* gene is located in a region syntenic to the *Tm-1* gene of tomato. The molecular marker, CaTm-int1, linked to *Cmr1*, was developed using the ortholog of *Tm-1*

EST (cacn2211), which has the most nucleotide similarity with *Tm-1*, was used to develop markers. A total of five introns were predicted in this EST sequence using the Intron Finder program of SGN (http://solgenomics.net/). Sequence analysis of the first intron revealed a polymorphic endonuclease recognition site, *Hinf*I, which was used to develop the CAPS marker CaTm-int1. CaTm-int1 mapped near TG31 on chromosome 2. However, CaTm-int1 was not polymorphic in the 'Bukang' F_2 population, but further analysis showed that the third intron sequence was polymorphic. This SNP was detected by high resolution melting (HRM) analysis, and the marker was named 'CaTm-in3HRM'. Co-segregation analysis of CaTm-int3HRM and the *Cmr1* gene revealed six recombinant individuals out of 309 individuals.

2.3.3 *Using Comparative Genetics: Markers Linked to the* L *Locus (Yang et al. 2009)*

2.3.3.1 Preliminary Research and Background

Comparative genetic analysis between pepper, potato, and tomato indicates that most resistance (R) genes are not randomly distributed in the genome. Clusters of R genes in corresponding regions of the three genomes often confer resistance to unrelated pathogen types (Grube et al. 2000). A comparative genetic map revealed that the *L* locus (resistant against Tobamovirus) of pepper, the *I2* locus (resistant against *Fusarium oxysporum*) of tomato and the *R3* locus (resistant against *Phytophthora infestans*) of potato are positioned in a syntenic R gene cluster at the end of the long arm of chromosome 11.

The *I2* locus of tomato was isolated by map-based cloning (Ori et al. 1997). Two main R gene clusters, SL8D and SL8E, were found in this locus. The *I2C-1* gene, in the main cluster SL8D, had been cloned and demonstrated to confer resistance against *Fusarium*. Comparative analysis was performed using molecular markers located in the genetic region around the SL8D and SL8E clusters to isolate the *R3* gene. *R3a*, a resistance gene against *Phytophthora*, was isolated in the corresponding region of SL8E, but not SL8D in which *I2C-1* was isolated, suggesting that molecular markers linked to the *L* gene could be developed and that, furthermore, the *L* gene could be isolated using genetic information from *R3a* in potato and *I2C-1* in tomato, because the former was isolated using genetic information of the latter (Fig. 2.21).

2.3.3.2 Plant Materials and Phenotype Analysis

F_2 *L*-segregating populations were constructed, and co-segregation analysis was performed for the linkage analysis between the *L* gene and polymorphic molecular markers. F_2 populations were derived by self-pollination of the F_1 commercial varieties Cupra (L^3/L^0), Special (L^4/L^1), and Myoung-Sung (L^4/L^1). The resistant

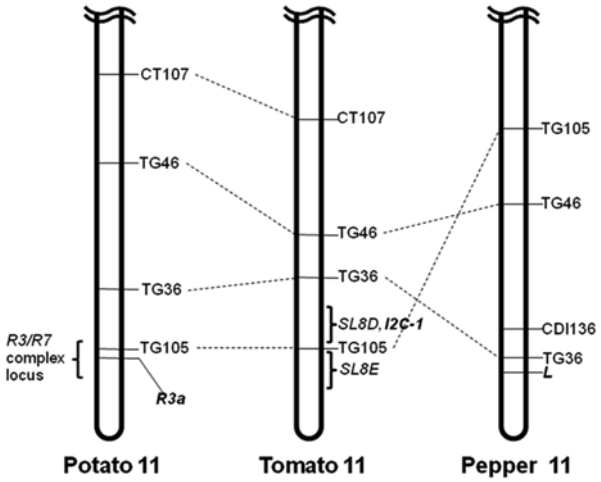

Fig. 2.21 Comparative map of the end of the long arm of chromosome 11 in potato, tomato, and pepper. The positions of *R3a* in potato, *I2C-1* in tomato, and *L* in pepper were conserved. Molecular marker analysis indicated that the syntenic relationship between potato and tomato covered most of the long arm, but genetic rearrangement occurred between tomato and pepper

phenotype was analyzed by inoculating the Tobamovirus P_0 pathotype to the Cupra F_2 population (L^3 is resistant but L^0 is susceptible to P_0 pathotype) and the $P_{1.2.3}$ pathotype to the Special and Myoung-Sung F_2 populations (L^4 is resistant but L^1 is susceptible to $P_{1.2.3}$ pathotype). Black local lesions were observed in resistant individuals, but symptoms such as mosaicism and necrosis of stem or leaves were observed in susceptible plants. The resistant-to-susceptible segregation ratio of the F_2 populations of Cupra, Special, and Myoung-Sung were 189:54, 537:109, and 504:341, respectively. The segregation ratio of the Cupra F_2 population satisfied the 3:1 Mendelian ratio of a single dominant resistance gene (P=0.3173), but the ratios of the Special and Myoung-Sung F_2 populations deviated significantly.

2.3.3.3 Development of Molecular Markers

To find the *I2* homolog in the pepper genome, hybridization of a 221,184 BAC library, derived from *Capsicum frutescens* BG2816 containing the L^2 allele, was performed using a probe based on the 3′ sequence of *I2C-1*. A total of 89 BACs with positive signal were selected and amplified by PCR using primers designed from genetic region around *R3a* to find the *R3a* homolog. Bands of 1.5 kb, 0.8–1.3 kb, and 1.2 kb in size were amplified from 37, 52, and 22 BAC clones using the R7-1, R7-2, and LRR primers, respectively. Among the 22 clones that amplified using LRR, bands were generated using both R7-1 and R7-2 from 12 BAC clones and using only one of the two primers from 10 BAC clones. These results suggested that

the 22 BAC clones may contain the *L* candidate gene, because they could be amplified using primers designed based on both *I2C-1* and *R3a* and presumably included genetic information of those two genes. The PCR products of the LRR amplifications were analyzed, and the sequences were classified into nine groups. A BAC clone from each group was selected for draft sequencing. SSR type markers were developed from the BAC draft sequence and mapped to a linkage map. The pepBAC082F3-5 and pepBAC060I2-H3 markers developed based on 082F3 and 060I2 BAC clones co-segregated perfectly with the TG36 marker located 5.2 cM away from the *L* locus, and the pepBAC337L21-1 marker designed based on 337L21 BAC mapped near TG105, which is located on chromosome 11 but far from the *L* locus.

To construct a contig, PCR was performed using the pepBAC060I2-E4 marker designed based on the 060I2 BAC sequence. Among the 89 BACs, bands were clearly amplified in 087H3, 158K24, 207E3, and 290J13. Primers were designed based on the end sequences of these four BACs. A contig of 13 BACs was constructed by the presence/absence of PCR amplicons, giving an overlap order using BAC end primers (Fig. 2.22). A total of 16 primer sets were designed from the contig sequence to develop an *L*-linked molecular marker. Bands of 420 bp and 422 bp were amplified using 087H3T7 and 207E13SP6 from *Capsicum chacoense* PI260429 (L^4) and *C. annuum* 'ECW (L^0)'. Sequence analysis revealed nine and five SNPs between these two accessions in the 087H3T7 and 207E13SP6 amplicons,

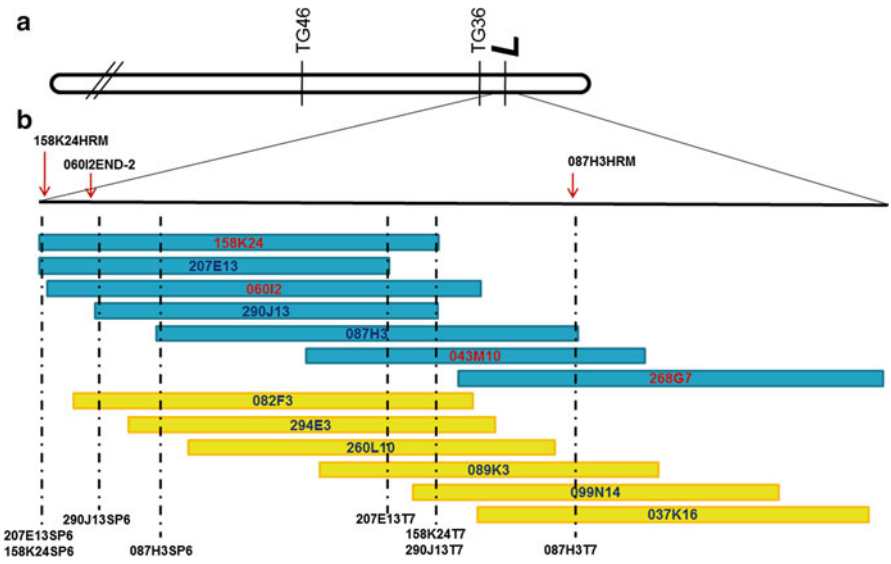

Fig. 2.22 A contig of 13 bacterial artificial chromosome (BAC) clones. The contig was constructed using a total of eight primers (listed below the *dotted lines*). Two co-dominant markers, 158K24HRM and 087G3HRM, and a dominant marker, 060I2END-2, were developed from BAC contig sequences

respectively. Primers containing one or two SNPs were designed for HRM analysis. The 087H03T7HRM marker redesigned based on the 087H03T7 BAC showed polymorphism (Fig. 2.22). A polymorphic marker, 158K24HRM, was developed in the same manner. The primer 060I2END-2, which was developed from 060I2, amplified a 700-bp band from ECW but not PI260429.

Co-segregation analysis used three polymorphic markers, 087H03T7HRM, 158K24HRM, and 060I2END-2, in three F_2 populations. The recombinant-to-total population size ratios of the three markers in Cupra, Special, and Myoung-Sung were 4/243, 5/361, and 11/858, respectively. Therefore, the genetic distances of these markers were calculated to be 0.8–1.2 cM from the *L* locus.

These case studies show different approaches to using genetic linkage mapping, bulked segregant analysis, comparative genetics, and physical mapping to develop molecular markers for key target loci.

References

Bonnet J, Danan S, Boudet C et al (2007) Are the polygenic architectures of resistance to *Phytophthora capsici* and *P. parasitica* independent in pepper? Theor Appl Genet 115:253–264

Grube RC, Radwanski ER, Jahn M (2000) Comparative genetics of disease resistance within the solanaceae. Genetics 155:873–887

Kang WH, Hoang NH, Yang HB et al (2010) Molecular mapping and characterization of a single dominant gene controlling CMV resistance in peppers (*Capsicum annuum* L.). Theor Appl Genet 120:1587–1596

Kim S, Hwang J, Kim G et al (2004) Development of markers linked to CMV resistant gene. Patent 10-2004-0086321, The Republic of Korea

Liu WY, Kang JH, Yang HB et al (2014) Combined use of bulked segregant analysis and microarrays reveals SNP markers pinpointing a major QTL for resistance to Phytophthora capsici in pepper. Theor Appl Genet 127:2503–2513

Ori N, Eshed Y, Paran I et al (1997) The I2C family from the wilt disease resistance locus *I2* belongs to the nucleotide binding, leucine-rich repeat superfamily of plant resistance genes. Plant Cell 9:521–532

Quirin EA, Ogundiwin EA, Prince JP et al (2005) Development of sequence characterized amplified region (SCAR) primers for the detection of *Phyto.5.2*, a major QTL for resistance to *Phytophthora capsici* Leon. in pepper. Theor Appl Genet 110:605–612

Thabuis A, Lefebvre V, Bernard G et al (2004) Phenotypic and molecular evaluation of a recurrent selection program for a polygenic resistance to *Phytophthora capsici* in pepper. Theor Appl Genet 109:342–351

Truong HTH, Kim KT, Kim DW et al (2012) Identification of isolate-specific resistance QTLs to phytophthora root rot using an intraspecific recombinant inbred line population of pepper (*Capsicum annuum*). Plant Pathol 61:48–56

Yang HB, Liu WY, Kang WH et al (2009) Development of SNP markers linked to the *L* locus in *Capsicum* spp. by a comparative genetic analysis. Mol Breed 24:433–446

http://www.springer.com/978-94-017-9995-9

Current Technologies in Plant Molecular Breeding
A Guide Book of Plant Molecular Breeding for Researchers
Koh, H.-J.; Kwon, S.-Y.; Thomson, M. (Eds.)
2015, XI, 352 p. 172 illus., 122 illus. in color., Hardcover
ISBN: 978-94-017-9995-9