
PBEA
PLANT BREEDING E-LEARNING IN AFRICA
Molecular Plant Breeding
Markers and Sequencing
START

Objectives

- Understand the importance of DNA sequence variation within species
- Learn the principles of sequencing
- Understand next generation (nextgen) sequencing
- Understand nextgen sequencing bioinformatics
- Understand prerequisites, prospects and limitations in using sequencing for genotyping
- Understand concept of imputation
- Understand RFLP, SSR and AFLP as examples of classical markers
- Understand SNP and INDEL markers basics
- Understand two basic applications of markers: fingerprinting and gene-tagging
- Understand underlying technologies of non-DNA marker systems and strengths and weaknesses of non-DNA versus DNA markers

Prerequisite

eModule on Mutation and Variation (Crop Genetics)

Fig. 1 A U.S. Food and Drug Administration microbiologist prepares DNA samples for gel electrophoresis analysis at the FDA lab in Atlanta. Photo by the U.S. Food and Drug Administration.

Introduction

Mutations are the ultimate source of all genetic variation. Mutations can occur at all levels of genetic organization, classified as **gene**, **chromosome** or **genome mutations**. Gene and chromosome mutations are discussed in this lesson. Gene mutations involve nucleotides or short DNA segments (one or few nucleotides substituted, inserted, or deleted). Chromosome mutations are large-scale chromosome alterations including deletions, insertions, inversions, and translocations. Genome mutations — involving changes in number of whole chromosomes or sets of chromosomes (see eModule 8 in Crop Genetics).

Genetic variation — dissimilarity between individuals attributable to differences in genotype — that is generated by mutations is acted upon by various evolutionary forces. Evolutionary processes that alter species and populations include selection, gene flow (migration), and genetic drift — whether plants are cultivated or wild. Evolution can be defined as a change in gene frequency over time. The way that plants evolve is dependent on both genetic characteristics and the environment that they face.

Introduction

Genetic variation results from differences in DNA sequences and, within a population, occurs when there is more than one allele present at a given locus. Changes in gene frequencies within populations caused by natural selection can lead to enhanced adaptation, while changes caused by human-directed selection can facilitate development of useful genetic variability and selection of superior genotypes. Selection is the differential reproduction of the products of recombination — both within and between chromosomes.

The basic tools used to characterize genetic variation within and between populations are called **genetic markers**. Markers can be visible traits, proteins, genes, DNA sequences, or RNA sequences, and can be genetically mapped to a particular chromosomal location. They can be used to track the inheritance of nearby genes to which they are closely linked. A marker may be part of a gene itself or more commonly in a chromosome segment close by a gene of interest. Markers are characteristically locus-specific and **polymorphic** (i.e., segregating) in the population under study, and also have easily observable phenotypes. Markers allow determination of alleles present in individuals or populations.

Introduction

Historically, plant breeders seeking sources of variability were constrained in choice of parental materials or **plant genetic resources** that were interfertile through the process of recombining locally adapted plant materials via sexual reproduction within closely related **gene pools** or just by evaluation and selection of particular desirable, existing genotypes to produce improved plant varieties. But a range of new techniques such as mutation induction, **genetic engineering** (transgenic or transformed plants), and in vitro methods (**somatic cell hybridization**, **tissue culture**, **doubled haploids**, **induced polyploids**) expand the source and scope of variability that can be used in crop improvement.

Our expanding understanding of the molecular basis of genetics has provided insights and technologies that further not only our basic understanding of genes and their regulation, but also provide additional tools for crop improvement. Molecular techniques enable breeders to generate genetic variability, transfer genes between unrelated species, move synthetic genes into crops, and make selections at the molecular, cellular, or tissue levels. Combining these laboratory techniques with conventional field approaches can shorten the time and reduce the costs for developing improved cultivars. The importance and application of molecular technologies are rapidly increasing.

Principles of Sequencing

Sequencing is the determination of the order of the nucleotides on a DNA molecule. A major milestone in plant biology was reached when the genome of Arabidopsis thaliana was published (The Arabidopsis Genome Initiative, 2000). Thereafter, the scientific community pursued the genomes of several crop plants used for feed and food.

Fig. 2 Arabidopsis thaliana in bloom. Photo by Alberto Salguero; licensed under CC-SA 3.0 via Wikimedia Commons.

Common name	Scientific name	Year
Potato	Solanum tuberosum	2011
Grape	Vitis vinifera	2007
Cucumber	Cucumis sativus	2009
Poplar	Populus trichocarpa	2006
Strawberry	Fragaria vesca	2010
Castor bean	Ricinus communis	2010
Apple	Malus x domestica	2010
Cannabis	Cannabis sativa	2011
Lotus	Lotus japonicus	2008
Soybean	Glycine max	2010
Pigeon pea	Cajanus cajan	2011
Chocolate	Theobroma cacao	2010
Papaya	Carica papaya	2008
Arabidopsis	Arabidopsis thaliana	2000
Arabidopsis	Arabidopsis lyrata	2011
Various	Brassica rapa	2011
Thellungiella	Thellungiella parvula	2011
Date palm	Phoenix dactylifera	2011
Rice	Oryza sativa L.	2002
Brachy	Brachypodium distachyon	2010
Maize	Zea mays	2009
Sorghum	Sorghum bicolor	2009
Moss	Physcomitrella patens	2008
Selaginella	Selaginella moellendorffii	2011

Principles of Sequencing

SANGER'S DIDEOXY DNA-SEQUENCING PROCEDURE

This procedure was developed by Fred Sanger in the 1970s. Sanger, along with Walter Gilbert, won the Nobel Prize in chemistry in 1980 for their sequencing developments. The method uses enzymatic reactions to incorporate specific terminators of DNA chain elongation called 2′,3′-dideoxynucleoside triphosphates (ddNTPs). The ddNTP molecules can be incorporated into the growing DNA chain through their 5′ triphosphate groups. However, because they lack a hydroxyl (OH) groups on the 3′-C of the sugar moiety, they cannot form a phosphodiester bond with deoxynucleotide triphosphates (dNTPs) during the sequencing reaction, resulting in termination of DNA chain elongation.

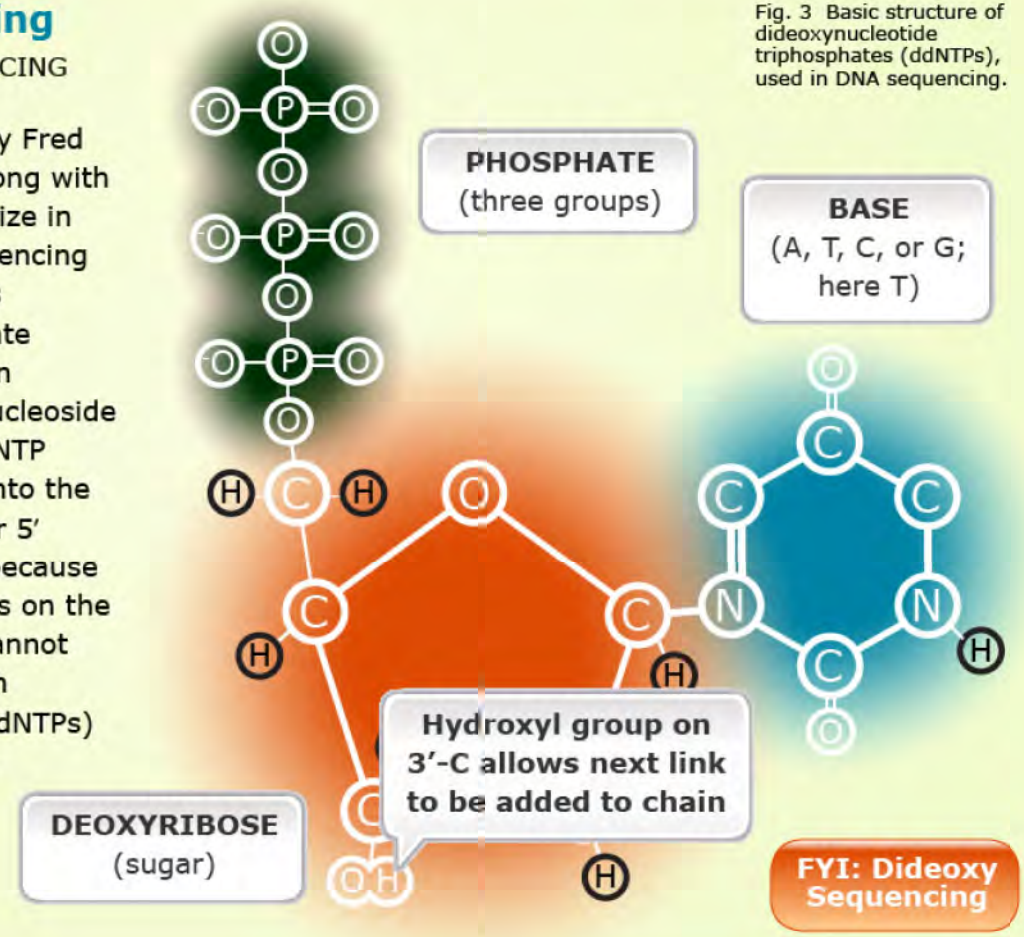


Fig. 3 Basic structure of dideoxynucleotide triphosphates (ddNTPs), used in DNA sequencing.

Principles of Sequencing

Plant name	Status	References	Plant name	Status	References
Arabidopsis	Published	Link	Brachypodium	Published	Link
Rice	Published	Link	Sugar beet	In-progress	Link
Maize	Published	Link	Flax	In-progress	Link
Sorghum	Published	Link	Cassava	In-progress	Link
Soybean	Published	Link	Peach	In-progress	Link
Potato	Published	Link	Common bean	In-progress	Link
Cucumber	Published	Link	Cacao	In-progress	Link
Apple	Published	Link	Sweet orange	In-progress	Link
Papaya	Published	Link	Sunflower	In-progress	Link
Medicago	Published	Link	Wheat	In-progress	Link
Grape	Published	Link	Barley	In-progress	Link
Poplar	Published	Link	Watermelon	In-progress	Link
Castor bean	Published	Link	Amborella	In-progress	Link
Pigeon Pea	Published	Link	Tomato	In-progress	Link
Strawberry	Published	Link	Peanut	Planned	Link
Date Palm	Published	Link			

Click here to see a video describing Sanger DNA sequencing.

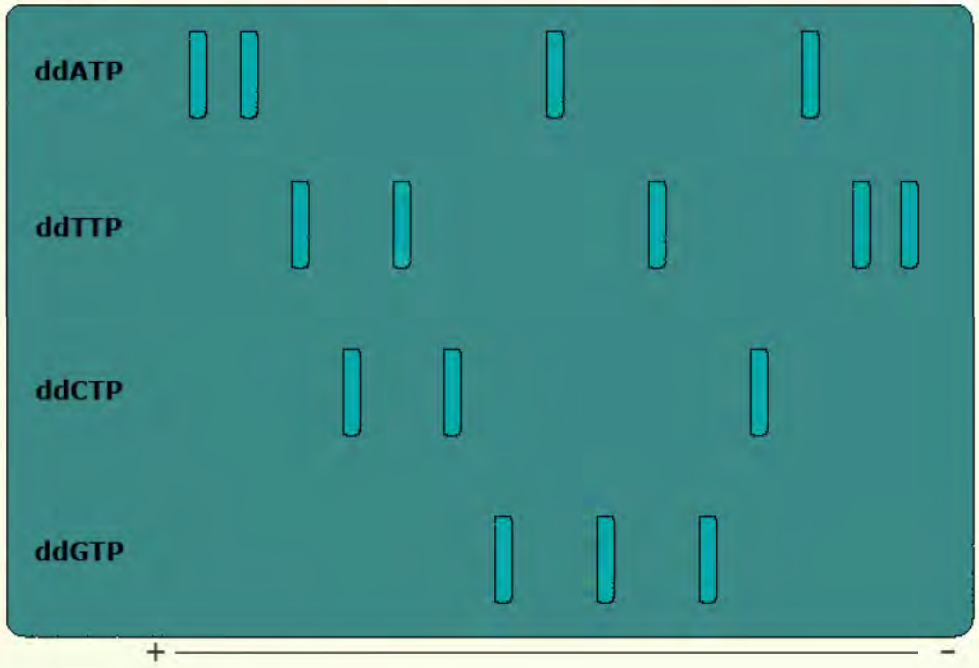
Table 1. Examples of plant species, whose genomes have been sequenced and published, or are in the process of being sequenced.

Updated overview over sequenced plant genomes:
http://www.ncbi.nlm.nih.gov/genome/browse/

TRY THIS!

Dideoxy Sequencing

The figure below shows gel electrophoresis sequencing results based on the dideoxy method. Examine the banding pattern and describe the order of the nucleotides of the original template strand.



Principles of Sequencing

MAXAM & GILBERT DNA-SEQUENCING PROCEDURE

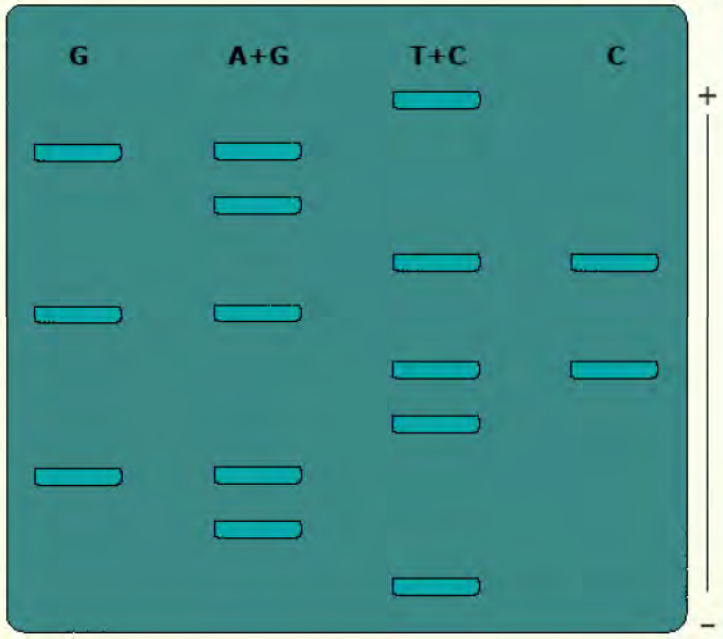
This procedure was developed by Allan Maxam and Walter Gilbert in 1977. The procedure is based on chemical degradation of DNA chains. In this procedure, a segment of DNA is labeled at one end with a radioactive label (^{32}P ATP). A solution containing the labeled DNA is distributed into four different tubes. A chemical that specifically destroys one or two of the four bases (G, A+G, C, C+T) in the DNA is added into each tube. Addition of the chemical piperidine to the DNA results in cleavage of the strand at the position of the modified base. The length of the cleaved fragments depends on the distance between the modified base and the labeled end of the DNA segment. The cleaved products of each of the four reactions (G, A+G, C, and C+T) will be evaluated by autoradiography and the banding pattern on film is scored to determine the DNA sequence.

FYI: Maxam & Gilbert Sequencing

TRY THIS!

Maxam and Gilbert DNA Sequencing

Determine the sequence of the analyzed strand and its complement based on the gel electrophoresis result shown below.



Analyzed strand:

5′– –3′

Complementary strand:

3′– –5′

More details (PDF)

Next Generation Sequencing (NGS)

Next generation sequencing is defined as a high-throughput sequencing method that combines parallel processes to produce millions of sequences at once. Several nextgen technologies are currently in use. The lesson focuses on the following technologies: pyrosequencing, Illumina, SOLiD, single molecule real time, and ion torrent sequencing.

Fig. 4 An Ion Torrent sequencing system. Photo by ThermoFisher Scientific.

Next Generation Sequencing (NGS)

PYROSEQUENCING OR 454 SEQUENCING

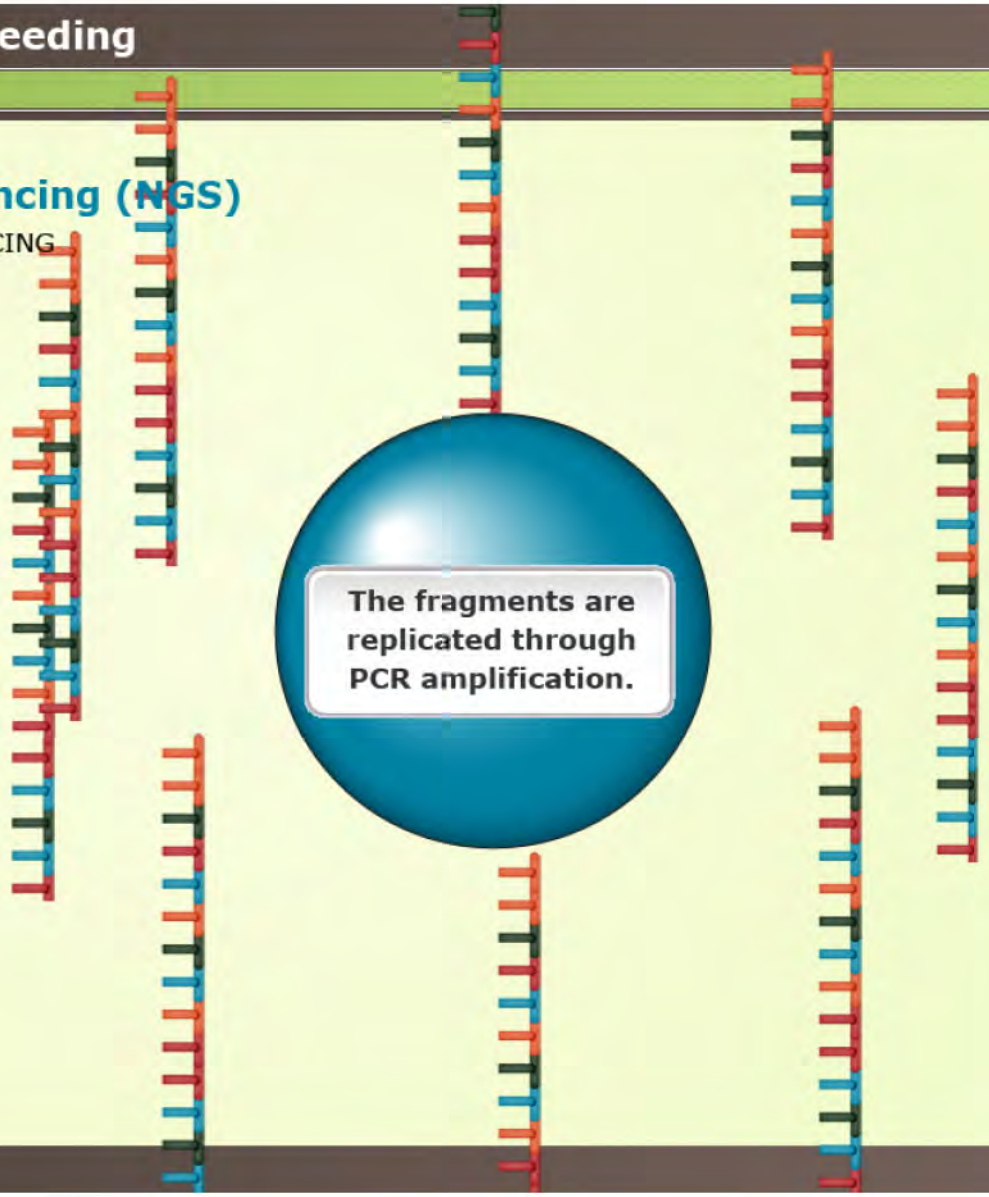
454 sequencing was developed by Roche and uses a procedure known as sequencing by synthesis, or pyrosequencing.

Fragments of the DNA to be sequenced are bound to a bead.

Next Generation Sequencing (NGS)

PYROSEQUENCING OR 454 SEQUENCING

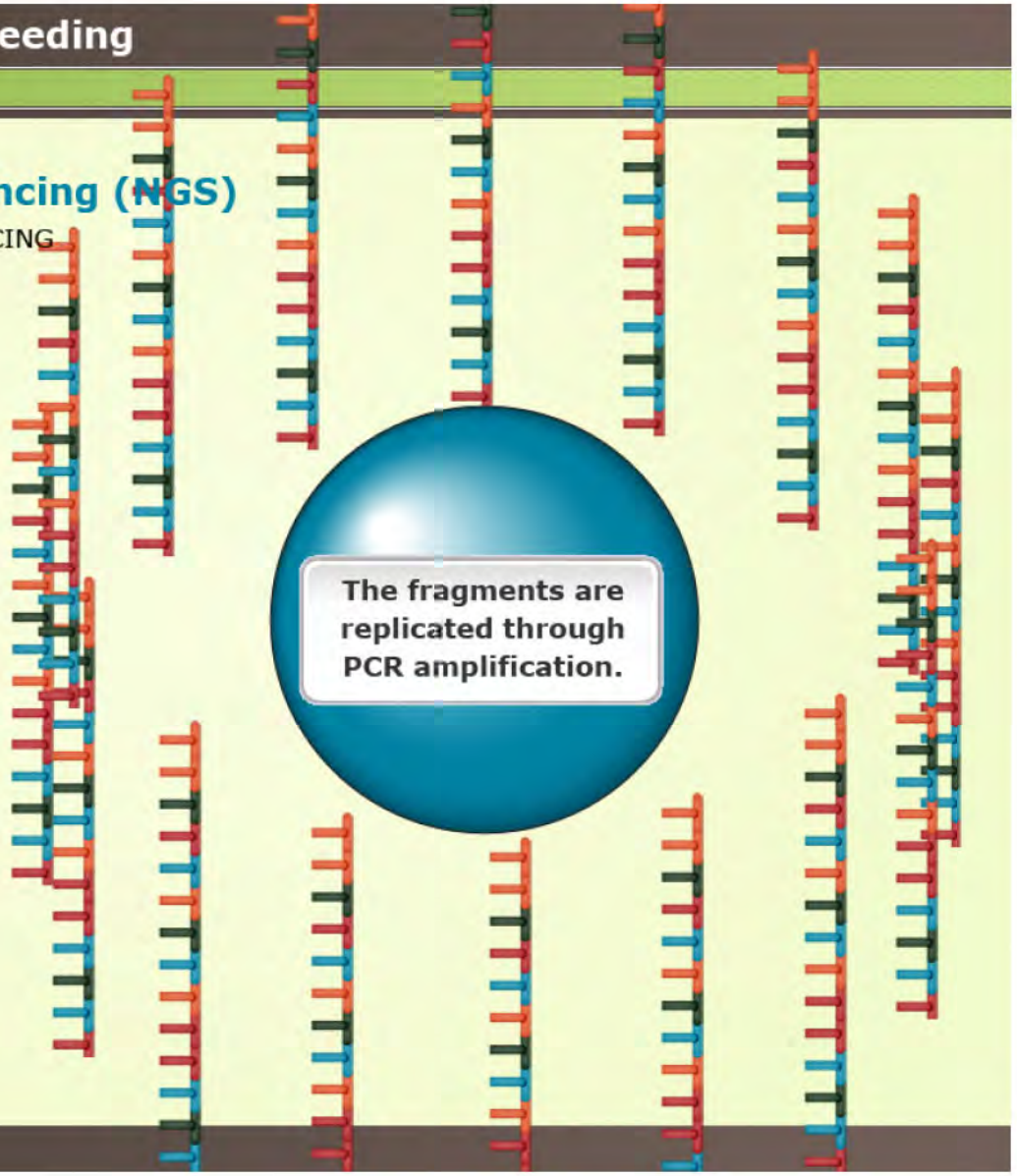
454 sequencing was developed by Roche and uses a procedure known as sequencing by synthesis, or pyrosequencing.



Next Generation Sequencing (NGS)

PYROSEQUENCING OR 454 SEQUENCING

454 sequencing was developed by Roche and uses a procedure known as sequencing by synthesis, or pyrosequencing.

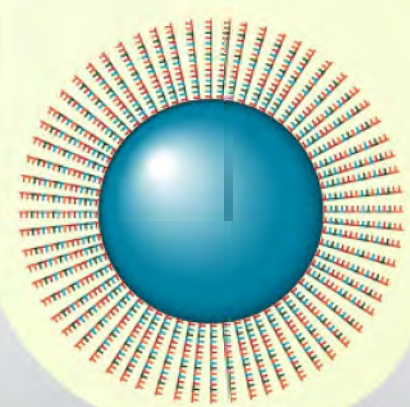


Next Generation Sequencing (NGS)

PYROSEQUENCING OR 454 SEQUENCING

454 sequencing was developed by Roche and uses a procedure known as sequencing by synthesis, or pyrosequencing.

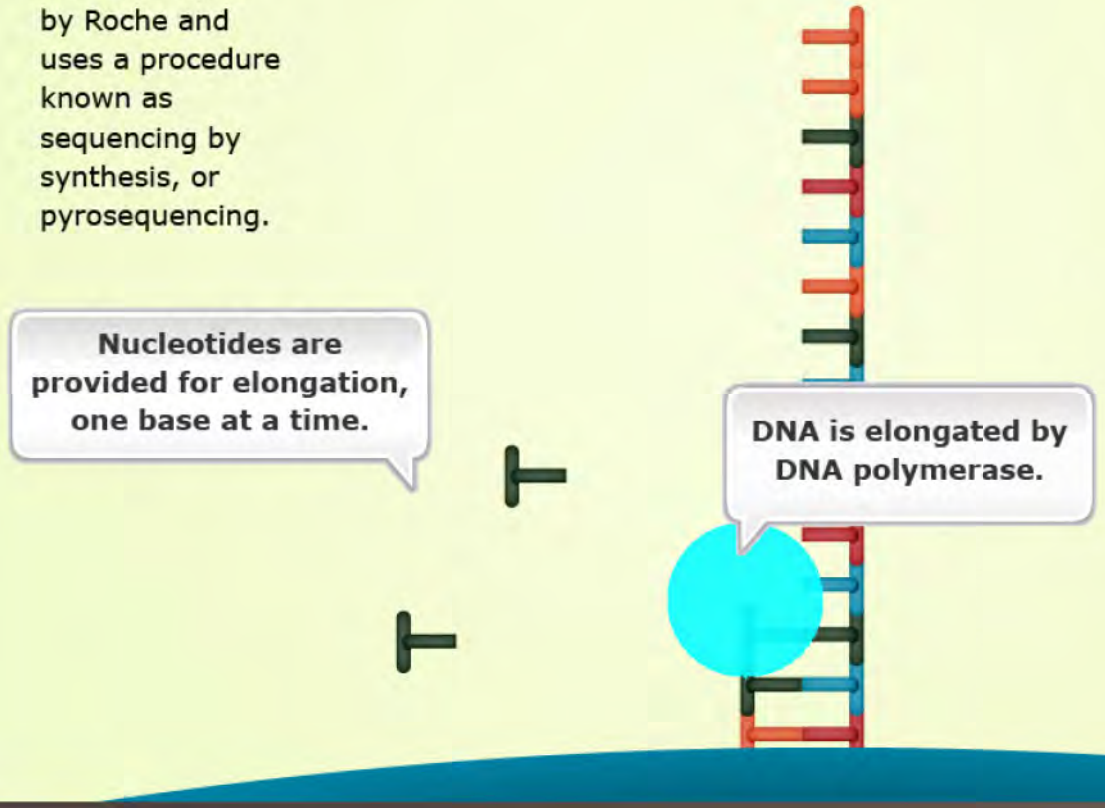
Each bead is placed in a well in preparation for chemical baths.



Next Generation Sequencing (NGS)

PYROSEQUENCING OR 454 SEQUENCING

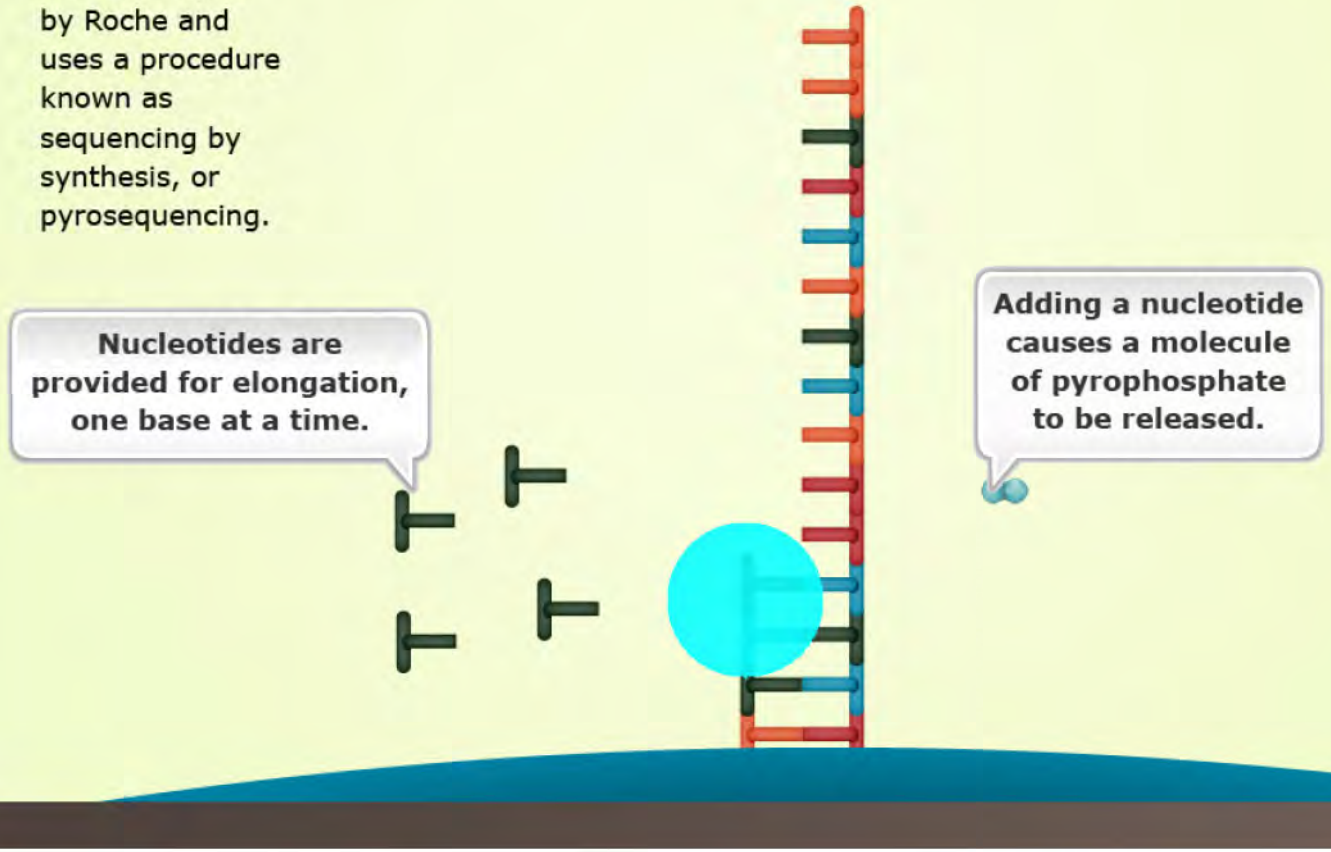
454 sequencing was developed by Roche and uses a procedure known as sequencing by synthesis, or pyrosequencing.



Next Generation Sequencing (NGS)

PYROSEQUENCING OR 454 SEQUENCING

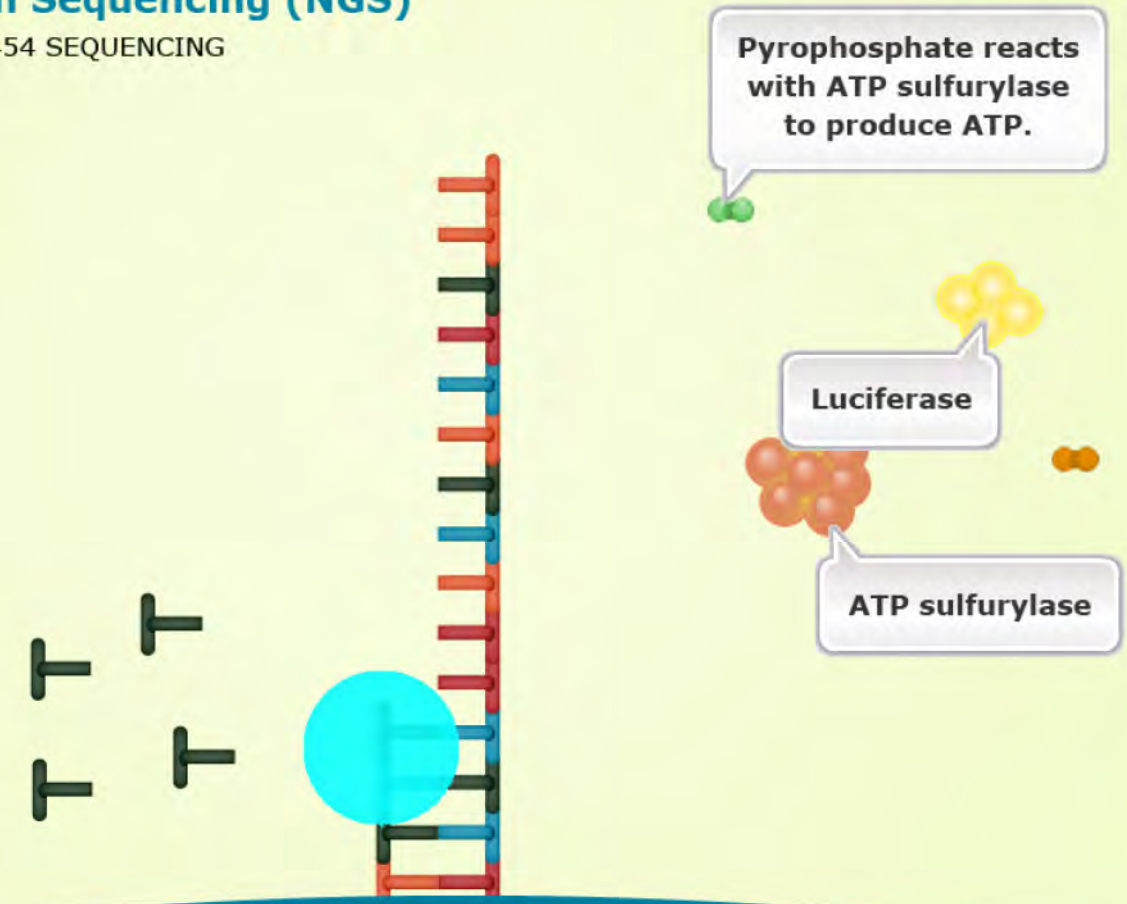
454 sequencing was developed by Roche and uses a procedure known as sequencing by synthesis, or pyrosequencing.



Next Generation Sequencing (NGS)

PYROSEQUENCING OR 454 SEQUENCING

454 sequencing was developed by Roche and uses a procedure known as sequencing by synthesis, or pyrosequencing.

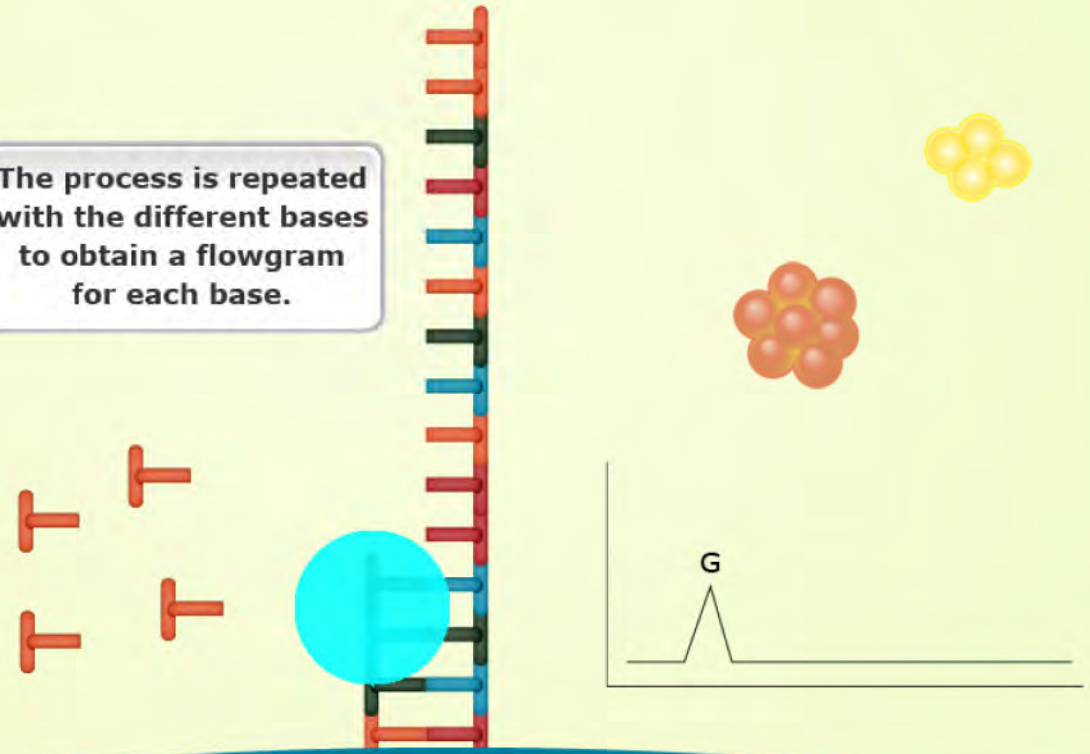


Next Generation Sequencing (NGS)

PYROSEQUENCING OR 454 SEQUENCING

454 sequencing was developed by Roche and uses a procedure known as sequencing by synthesis, or pyrosequencing.

The process is repeated with the different bases to obtain a flowgram for each base.

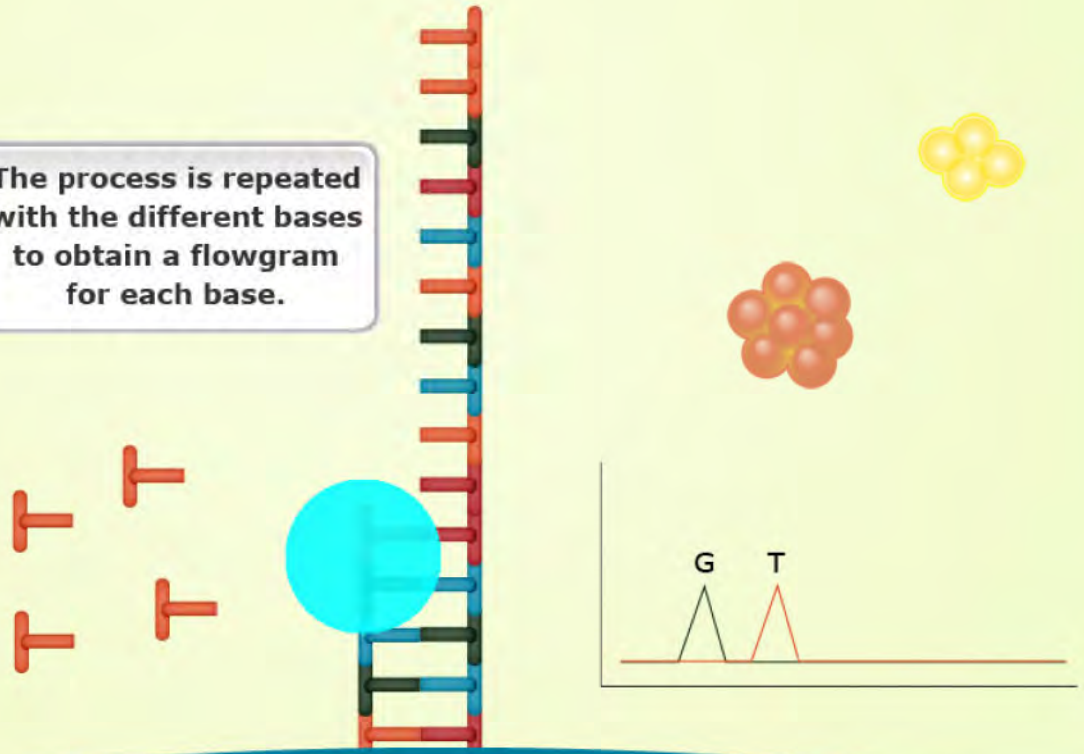


Next Generation Sequencing (NGS)

PYROSEQUENCING OR 454 SEQUENCING

454 sequencing was developed by Roche and uses a procedure known as sequencing by synthesis, or pyrosequencing.

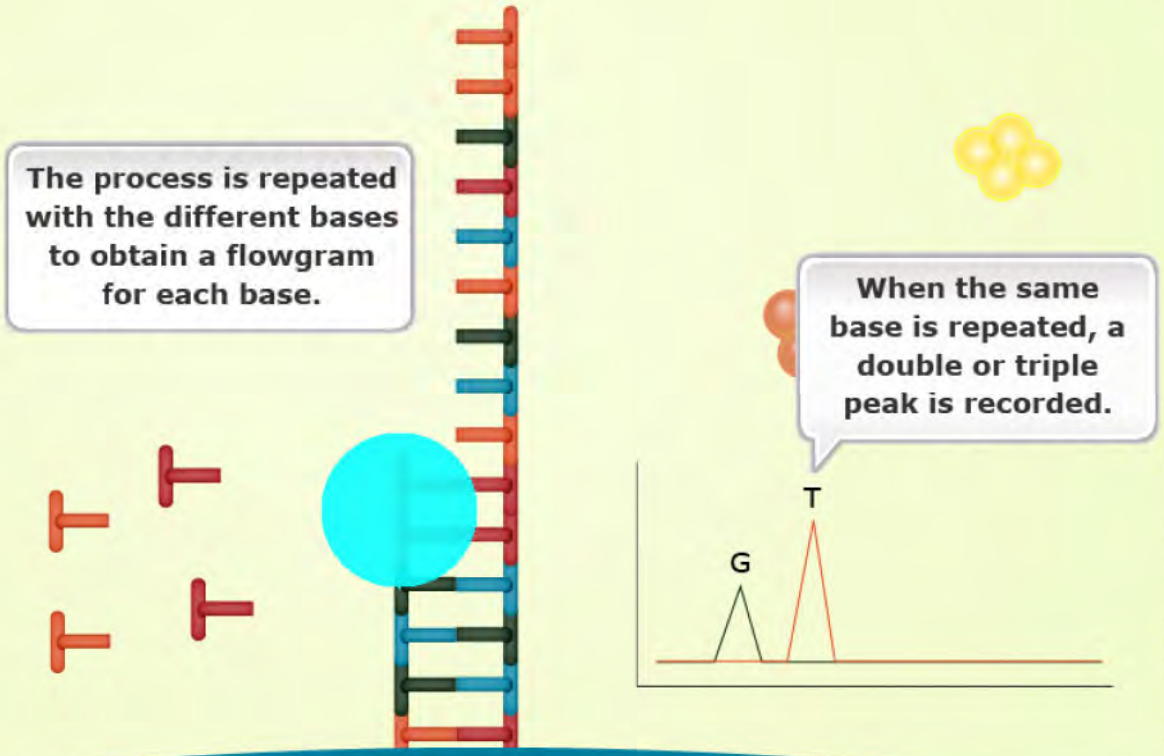
The process is repeated with the different bases to obtain a flowgram for each base.



Next Generation Sequencing (NGS)

PYROSEQUENCING OR 454 SEQUENCING

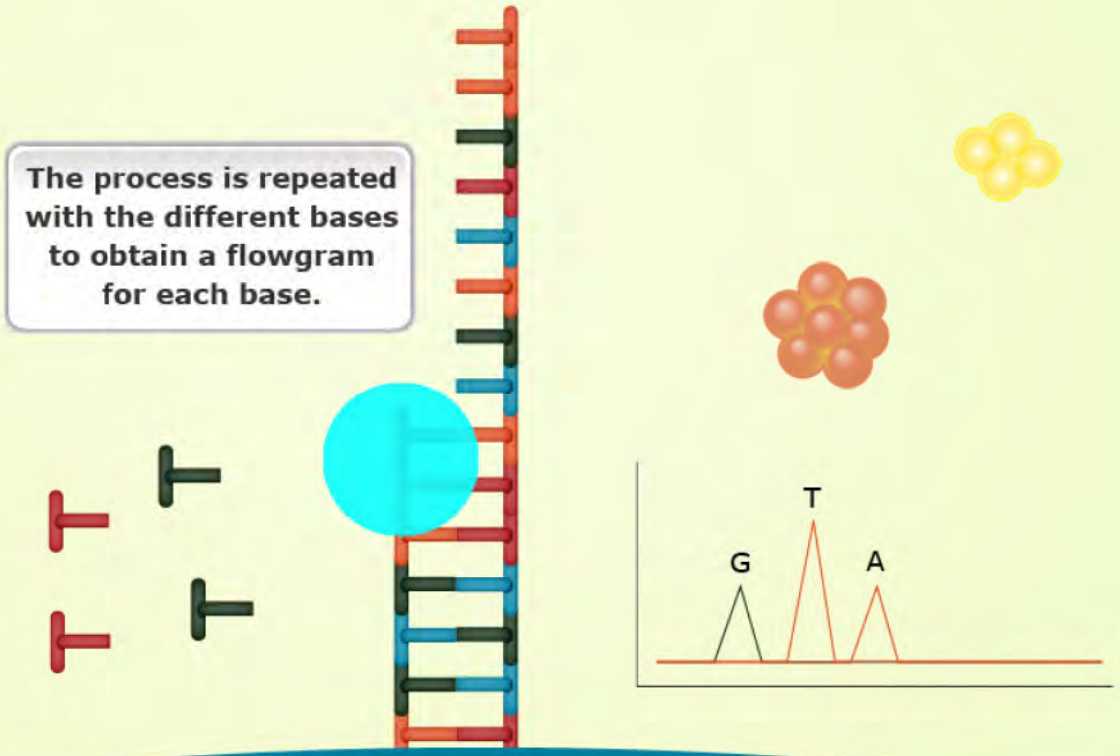
454 sequencing was developed by Roche and uses a procedure known as sequencing by synthesis, or pyrosequencing.



Next Generation Sequencing (NGS)

PYROSEQUENCING OR 454 SEQUENCING

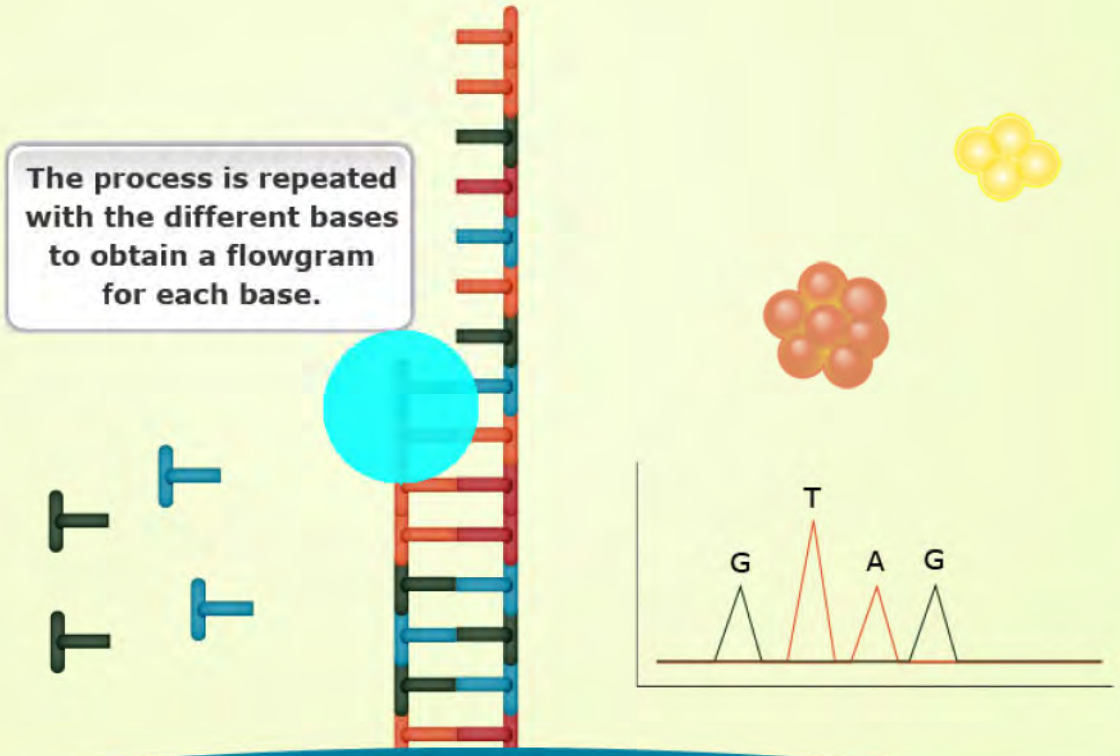
454 sequencing was developed by Roche and uses a procedure known as sequencing by synthesis, or pyrosequencing.



Next Generation Sequencing (NGS)

PYROSEQUENCING OR 454 SEQUENCING

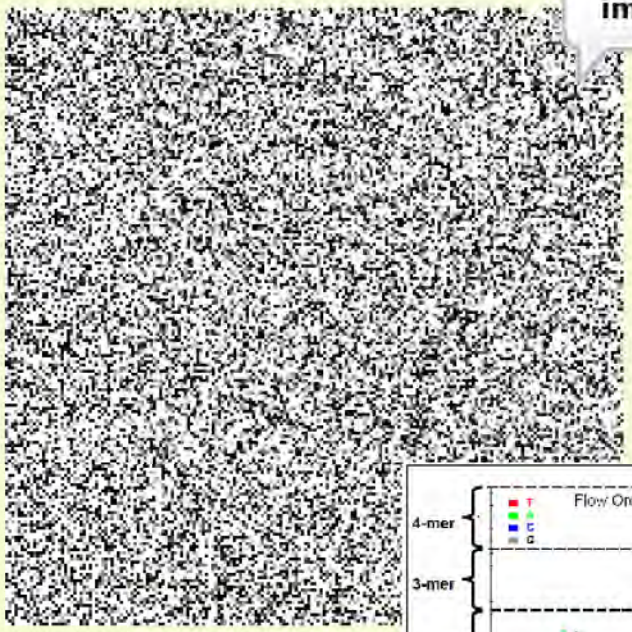
454 sequencing was developed by Roche and uses a procedure known as sequencing by synthesis, or pyrosequencing.



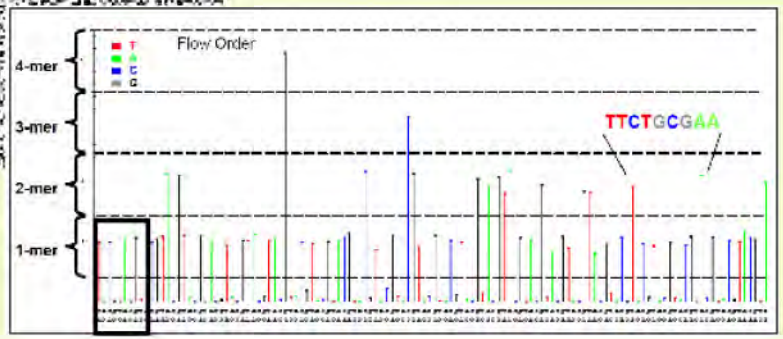
Next Generation Sequencing (NGS)

PYROSEQUENCING OR 454 SEQUENCING

454 sequencing was developed by Roche and uses a procedure known as sequencing by synthesis, or pyrosequencing.



A computer system decodes the signal images captured by a CCD array and converts them into a dataflow diagram which indicates the order of bases on the DNA strand.

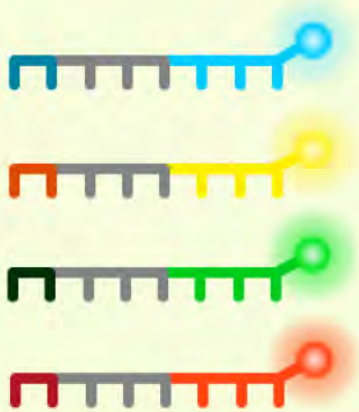


Next Generation Sequencing (NGS)

SOLiD

The SOLiD system was developed by Life Technologies and is based on a technique of oligonucleotide ligation and detection. Like pyrosequencing, SOLiD sequencing begins with fragmented DNA on an agarose bead.

A series of probes are ligated to the template strand. Each probe has four main components.



Interrogation bases

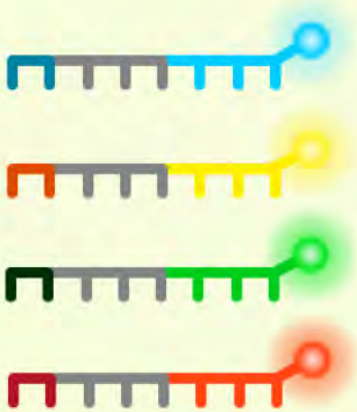
DNA template strand

Next Generation Sequencing (NGS)

SOLiD

The SOLiD system was developed by Life Technologies and is based on a technique of oligonucleotide ligation and detection. Like pyrosequencing, SOLiD sequencing begins with fragmented DNA on an agarose bead.

A series of probes are ligated to the template strand. Each probe has four main components.



Degenerate bases

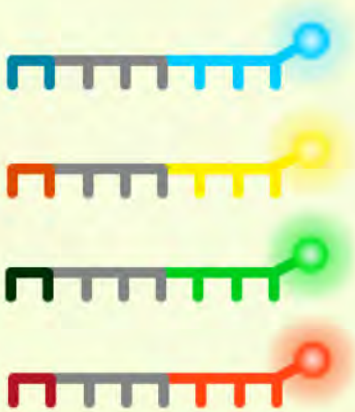
DNA template strand

Next Generation Sequencing (NGS)

SOLiD

The SOLiD system was developed by Life Technologies and is based on a technique of oligonucleotide ligation and detection. Like pyrosequencing, SOLiD sequencing begins with fragmented DNA on an agarose bead.

A series of probes are ligated to the template strand. Each probe has four main components.



Universal bases

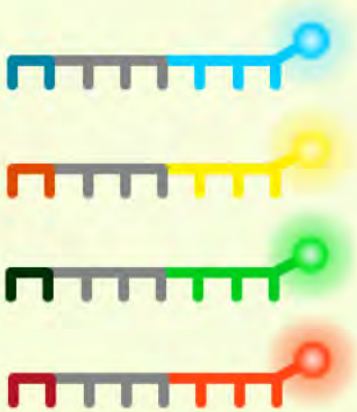
DNA template strand

Next Generation Sequencing (NGS)

SOLiD

The SOLiD system was developed by Life Technologies and is based on a technique of oligonucleotide ligation and detection. Like pyrosequencing, SOLiD sequencing begins with fragmented DNA on an agarose bead.

A series of probes are ligated to the template strand. Each probe has four main components.



Fluorescent dye

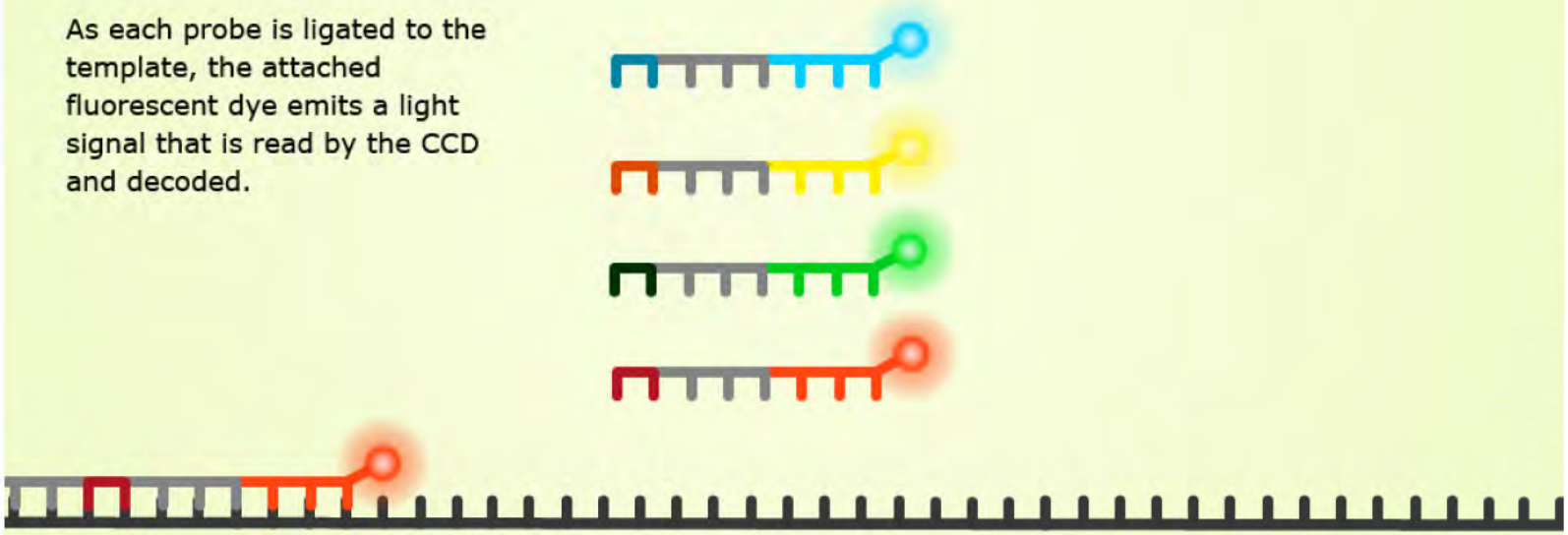
DNA template strand

Next Generation Sequencing (NGS)

SOLiD

The SOLiD system was developed by Life Technologies and is based on a technique of oligonucleotide ligation and detection. Like pyrosequencing, SOLiD sequencing begins with fragmented DNA on an agarose bead.

As each probe is ligated to the template, the attached fluorescent dye emits a light signal that is read by the CCD and decoded.



DNA template strand

Next Generation Sequencing (NGS)

ILLUMINA

The Illumina system uses a terminator-based method to detect single bases as they are incorporated into a growing DNA strand.

DNA is broken up into fragments...

Next Generation Sequencing (NGS)

ILLUMINA

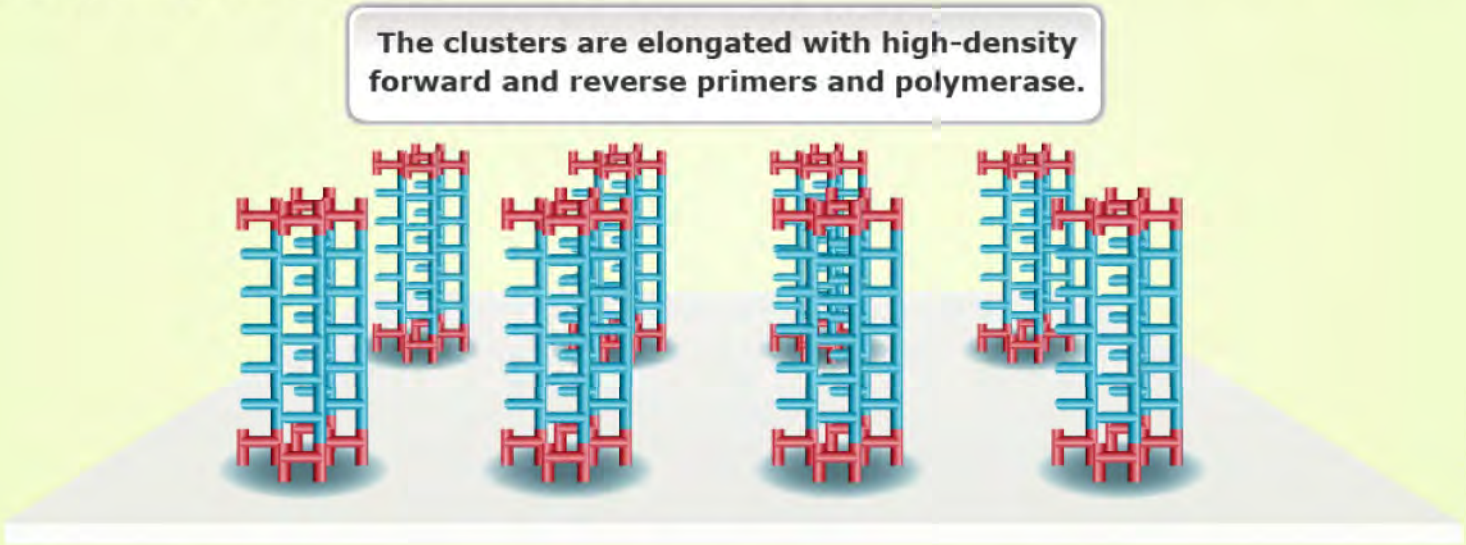
The Illumina system uses a terminator-based method to detect single bases as they are incorporated into a growing DNA strand.

The fragments are attached to a slide and amplified into clusters.

Next Generation Sequencing (NGS)

ILLUMINA

The Illumina system uses a terminator-based method to detect single bases as they are incorporated into a growing DNA strand.



Third-Generation Sequencing

SINGLE-MOLECULE REAL TIME (SMRT) SEQUENCING

The SMRT system was developed by Pacific Biosciences and uses a polymerase based approach to sequence single DNA molecules in real-time.

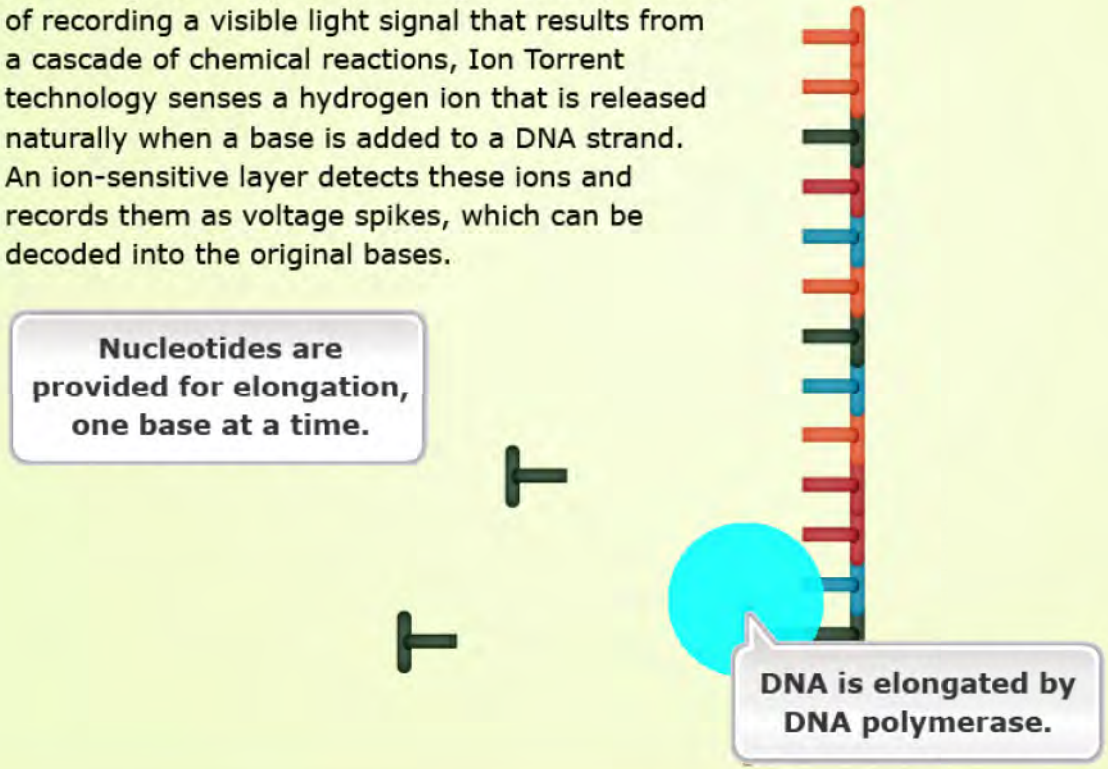
The technique works similarly to 454 pyrosequencing, but it uses a luminous dye attached to the phosphate chain of each nucleotide.

The phosphate chain is sheared off during normal DNA synthesis, which causes the dye to give a light signal that is captured by the system.

Third-Generation Sequencing

ION TORRENT SEQUENCING

Ion Torrent sequencing works similarly to NGS technologies like pyrosequencing. However, instead of recording a visible light signal that results from a cascade of chemical reactions, Ion Torrent technology senses a hydrogen ion that is released naturally when a base is added to a DNA strand. An ion-sensitive layer detects these ions and records them as voltage spikes, which can be decoded into the original bases.



Sequence Alignment

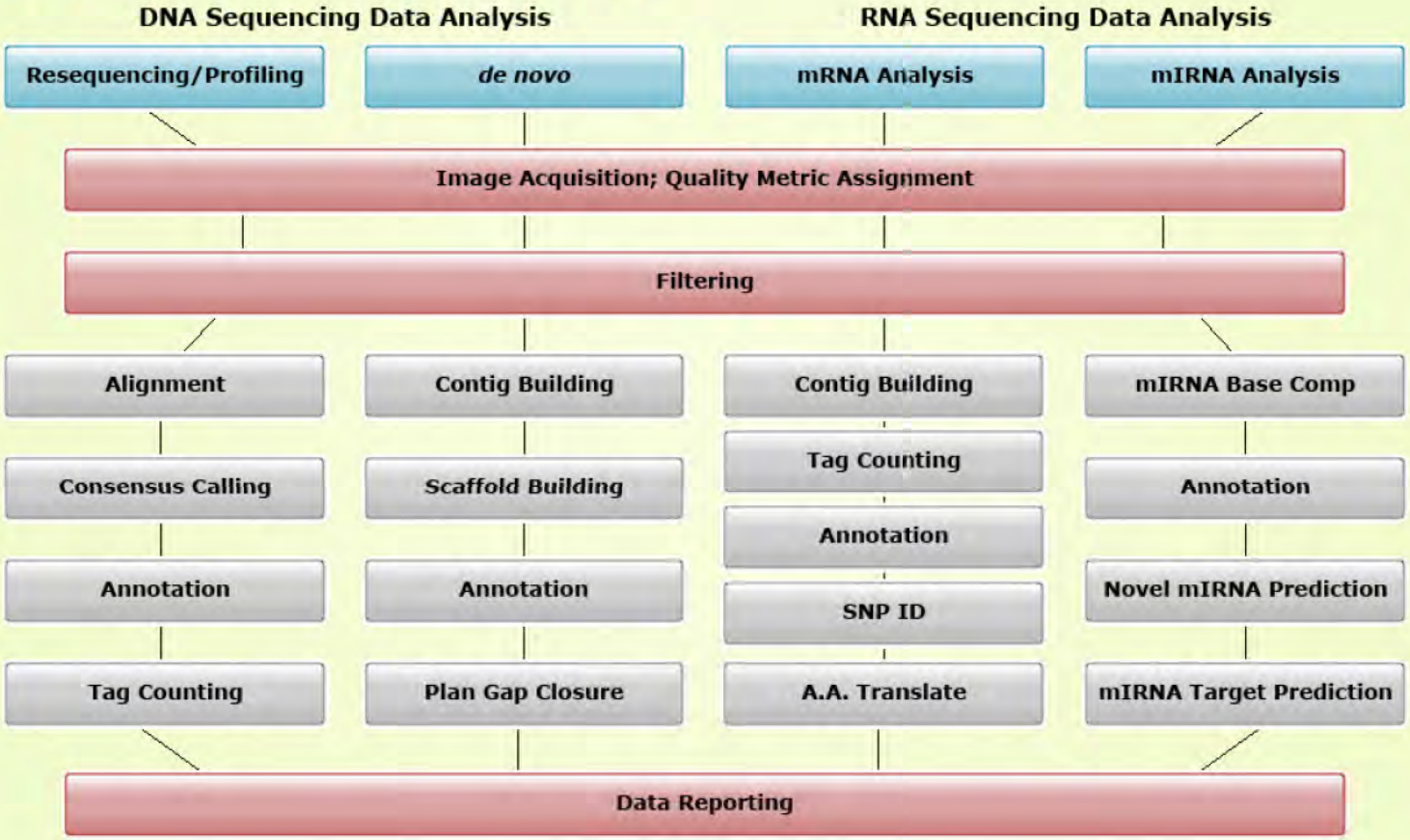
In general, nextgen technologies result in very large numbers of reads that are shorter than those produced using capillary electrophoresis. Therefore, nextgen sequencing requires more robust algorithms to assemble the large quantity of data generated. Along with the increased output, the challenge is to manage and track sequence runs, and automating downstream data analyses. Several academic and private institutions provide core services for nextgen sequencing. Among them are Beckman Coulter Genomics, Massachusetts General Hospital and the Office of Biotechnology at Iowa State University.

BECKMAN COULTER GENOMICS

MGH next Sequencing Core generation

Office of Biotechnology Iowa State University

General Bioinformatics Workflow



Third-Generation Sequencing

Table 2: Information regarding the various HTP sequencing platforms including sequence length, sequences per run and estimated error rate. Sources: (Glenn et al. 2011; Ozsolak 2012), www.illumina.com, Ion Torrent Application Note Spring 2011, www.appliedbiosystems.com, www.my454.com, www.pacificbiosciences.com.

Company	Platform	Read length (bp)	Reads per run (estimated)	Estimated error rate
Roche 454	GS FLX Titanium XL+ GS Junior	400-1000	100,000	0.4-1.5%
Illumina	HiSeq 2000 HiSeq 1000 Genome Analyzer IIX Genome Analyzer IIE iScanSQ	2 x 36-151	3.4 million to 6 billion	0.5-2%
Life Sciences	5500 SOLiD 5500xl SOLiD	2 x 35-75	700 million to 1 billion	0.06-0.2%
Ion Torrent	Ion Torrent 314 Ion Torrent 318 Ion Proton Ion Proton II	200	800 million	1-3%
Helicos Biosciences	HeliScope	55	100,000 to 8 million	3-5%
Pacific Biosciences	PacBio RS	700-6000	10,000	13-15%

Suggested References

Metzker (2010) Sequencing technologies the next generation. Nature Reviews Genetics. 11:31-46 is ideal for covering the Objective 3.
http://www.nature.com/nrg/journal/v11/n1/pdf/nrg2626.pdf

Eid et al. 2009. Real-Time DNA sequencing from single polymerase molecules. Science 323: 133-138 http://www.sciencemag.org/content/323/5910/133.full.pdf

Rothberg et al. 2011. An integrated semiconductor device enabling non-optical genome sequencing. Nature. 475:348-352.
http://www.nature.com/nature/journal/v475/n7356/pdf/nature10242.pdf

Shendure, J., and H, Ji.2008. Next-generation DNA sequencing. Nature Biotechnology 26:1135-1145. http://www.nature.com/nbt/journal/v26/n10/pdf/nbt1486.pdf

DiGuistini et al. 2009. De novo genome sequence assembly of a filamentous fungus using Sanger, 454 and Illumina sequence data. Genome Biology 10:R94
http://www.biomedcentral.com/content/pdf/gb-2009-10-9-r94.pdf

Munroe, D., and T. J. R. Harris. 2010. Third-generation sequencing fireworks at Marco Island. Nature Biotechnol 28:426-428
http://www.nature.com/nbt/journal/v28/n5/pdf/nbt0510-426.pdf

Suggested References

Hawkins et al. (2010) Next-generation genomics: an integrative approach. Nature Review Genetics. 11:476-486. http://www.nature.com/nrg/journal/v11/n7/pdf/nrg2795.pdf

Schneeberger, K., and D. Weigel. 2011. Fast-forward genetics enabled by new sequencing technologies. Trends Plant Sci. 16:282-8. http://ac.els-cdn.com/S136013851100029X/1-s2.0-S136013851100029X-main.pdf?_tid=69c29f8b83bfc51c21438db4d67728a3&acdnat=1334590816_547510fc0545ec6df921fc0017404666.

Schneeberger et al. 2009. SHOREmap: simultaneous mapping and mutation identification by deep sequencing. Nat Methods. 6:550-551. http://www.nature.com/nmeth/journal/v6/n8/pdf/nmeth0809-550.pdf

Genotyping by Sequencing

Next-generation sequencing has made it possible to sequence entire plant genomes in much shorter time and at a lower cost than using the approaches based on Sanger dideoxy sequencing (Glenn, 2011). Sequencing of multiple related genomes using NGS technologies can be done to sample genetic diversity within and between germplasm. However, even with NGS technologies, species with large complex genomes are a challenge to sequence. To address this challenge, genotyping-by-sequencing (GBS) was developed as a tool for association studies and genomics-assisted breeding for various crops species, including those with complex genomes.

Genotyping by Sequencing

GBS uses restriction enzymes in combination with multiplex sequencing to reduce genome complexity sequencing cost (Fig. 5).

Library Construction

1. Adapters are placed on 96-well plate

   DNA samples added and dried

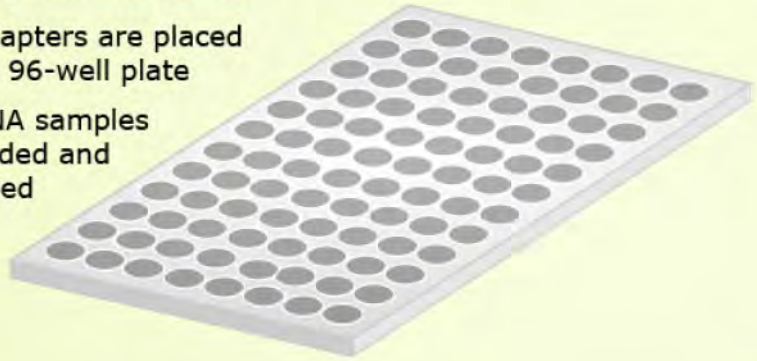


Fig. 5 Genotyping-by-sequencing in plants. Library construction involves plating the DNA and adapter pair, digestion with a restriction enzyme (or two), and ligation of adapters to the ends of DNA fragments. Samples are pooled and cleaned up before PCR. The PCR products are also cleaned up and evaluated for quality. Adapted from Elshire et al., 2011.

Genotyping by Sequencing

SEQUENCE BARCODING

NGS technologies can produce more than 1 billion base pairs in a single sequencing run. A challenge is, to use this enormous capacity for multiple DNA samples, for which only a fraction of the 1 billion bp sequence information is required. Barcoding enables to label sequences originating from a particular sample, and to pool barcoded DNA in a single sequencing run. Barcodes in the context of DNA sequencing are short, unique sequences of DNA added to samples to be pooled, then processed and sequenced in parallel (Fig. 6). The sequence produced from the barcoded samples contains information to determine its origin. By barcoding the DNA, base-by-base error rate and array-to-array or day-to-day variability are reduced.

Adapter and sequencing primer design

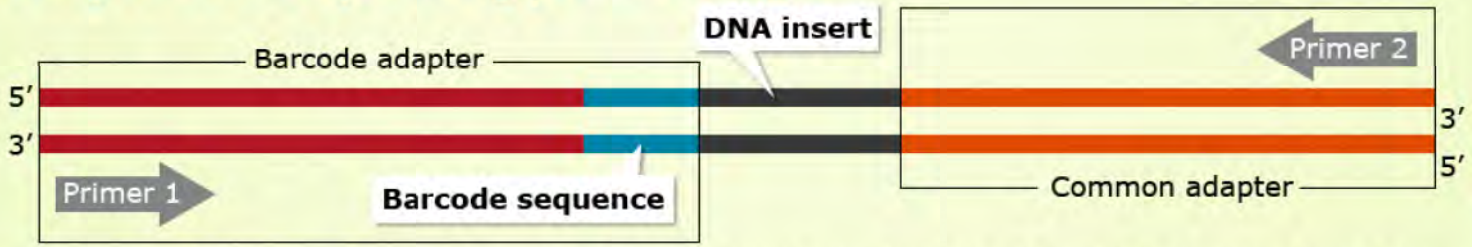


Fig. 6 Genotyping-by-sequencing in plants. A barcode adapter and a common adapter flank the DNA insert to be sequenced. Primers 1 and 2 bind specific sequences on the 3′ ends of the barcode and common adapters, respectively. Adapted from Elshire et al., 2011.

Genotyping by Sequencing

SEQUENCE BARCODING

Fig. 7 Preparation of barcoded libraries. Specific regions of genomes from multiple individuals are amplified by PCR. The PCR amplicons are pooled together and end-modified with barcoded adapters. The barcoded amplicons (referred to as an indexed library) are pooled together and sequenced by NGS. Adapted from Craig et al., 2008.

Genotyping by Sequencing

DEVELOPMENT OF HAPLOTYPE MAPS

Genome-wide association studies (GWAS) require both phenotypic and genotypic data from multiple individuals. The concept of GWAS will be covered in eModule4. Thus, GBS can be used to develop genotypic data for the construction of **haplotype** maps (HapMap) for GWAS. An example of the use of GBS for GWAS is from the work of Huang et al. (2010) describing the sequencing of 517 rice genomes using the Ilumina technology to generate about one-fold sequence coverage per genotype. The data generated by Huang et al. (2010) were used to construct a HapMap for GWAS for several agronomic traits in rice.

Data Imputation

The process of DNA sequencing is not free of error. Also, depending on the NGS system used, the length of base pair reads will be variable. Errors in sequencing and the length of the reads obtained by NGS may result in missing genotypes, thus affecting the quality of data.

The concern of missing data arises in almost all statistical analyses. This is actually what Huang and coworkers (Huang et al. 2010) encountered. Importantly, Huang and coworkers understood that **linkage disequilibrium** (LD) and the nonrandom correlation among allelic variants is extensive in rice. This meant that they could infer missing genotypes (missing data) with high confidence using data imputation (Marchini et al. 2007). Imputation is a statistical term describing substitution of a value for missing data.

The following steps summarize the approach used by Huang and coworkers to assign values to missing genotypes.

Fig. 8 Rice plants. Photo by IRRI Images. Licensed under Creative Commons Attribution 2.0 via Wikimedia Commons.

Data Imputation

STEP 1: SNP IDENTIFICATION AND ANNOTATION

Single-base pair genotypes of 520 individuals obtained by Illumina sequencing were integrated to screen for single-nucleotide polymorphisms (SNPs) across the genome. Candidate SNPs were identified by comparing Illumina sequence data with the rice reference genome.

Imputation program

Multipoint method (PDF)

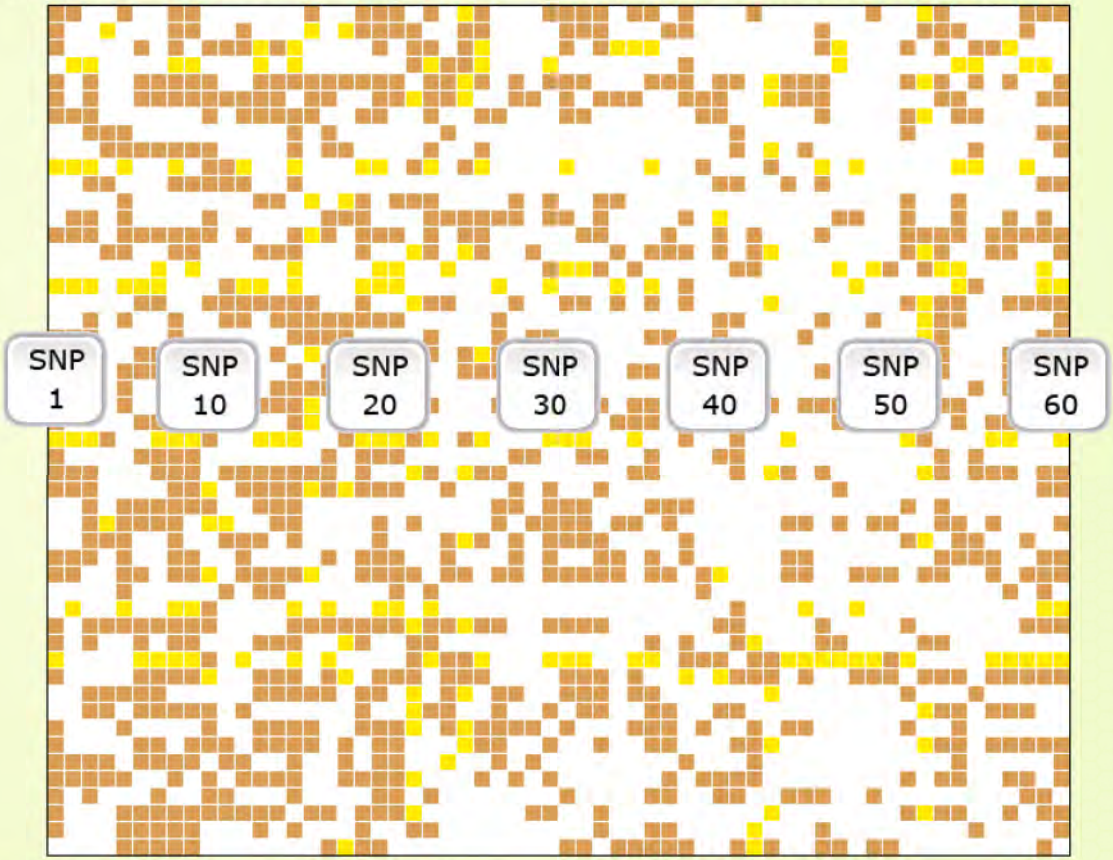
Data Imputation

STEP 2: DATA IMPUTATION TO ASSIGN GENOTYPES

A genotype matrix of 60 SNPs on the fourth chromosome of indica landraces obtained by sequencing.

Brown represents major alleles, gold stands for minor alleles, and white is for missing data.

Fig. 9 Transformation of missing genotypic values by imputation. Adapted from Huang et al., 2010.



Data Imputation

STEP 2: DATA IMPUTATION TO ASSIGN GENOTYPES

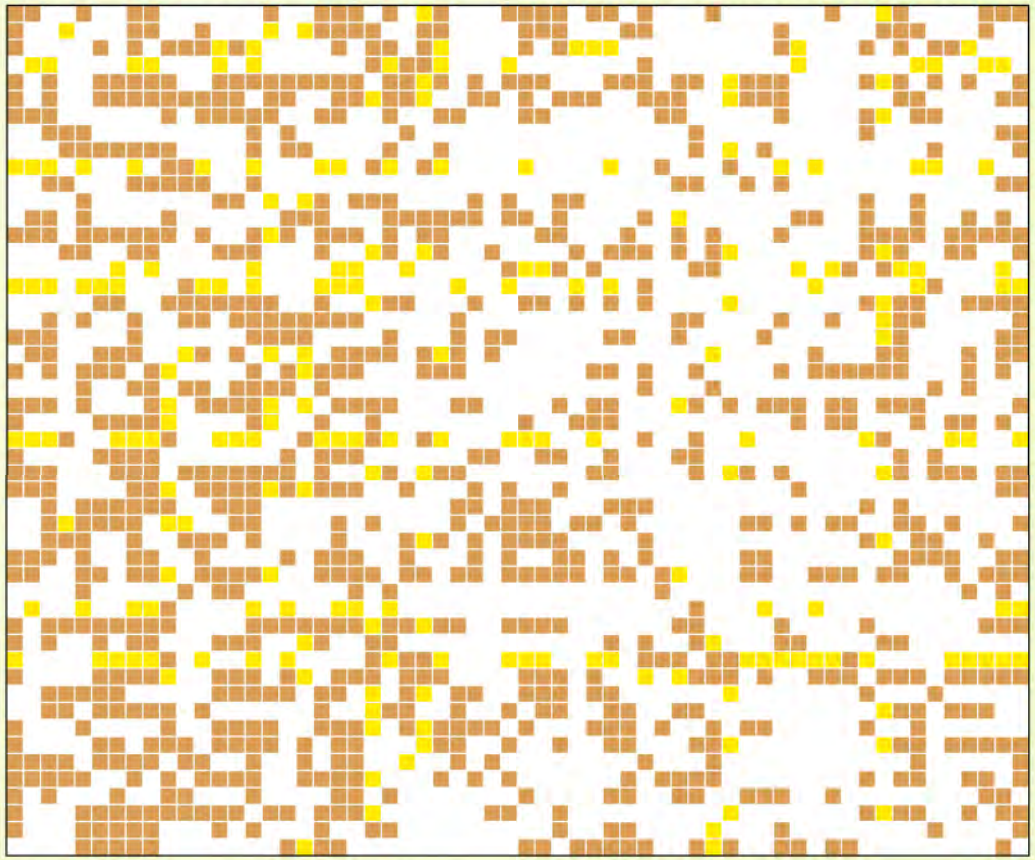
A genotype matrix of 60 SNPs on the fourth chromosome of indica landraces obtained by sequencing.

Brown represents major alleles, gold stands for minor alleles, and white is for missing data.

SNP sites are arranged horizontally, ordered according to chromosomal location.

Land races are arranged vertically. Each square in the matrix stands for a SNP genotype of a landrace.

Fig. 9 Transformation of missing genotypic values by imputation. Adapted from Huang et al., 2010.

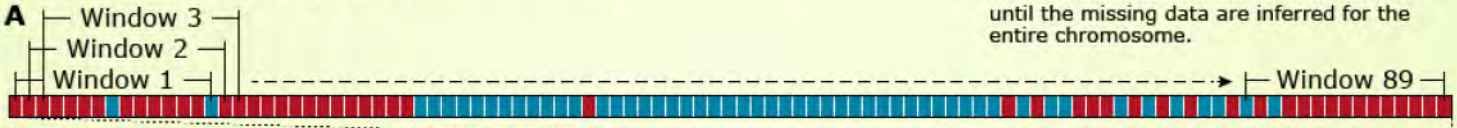


Data Imputation

THE SLIDING WINDOW APPROACH

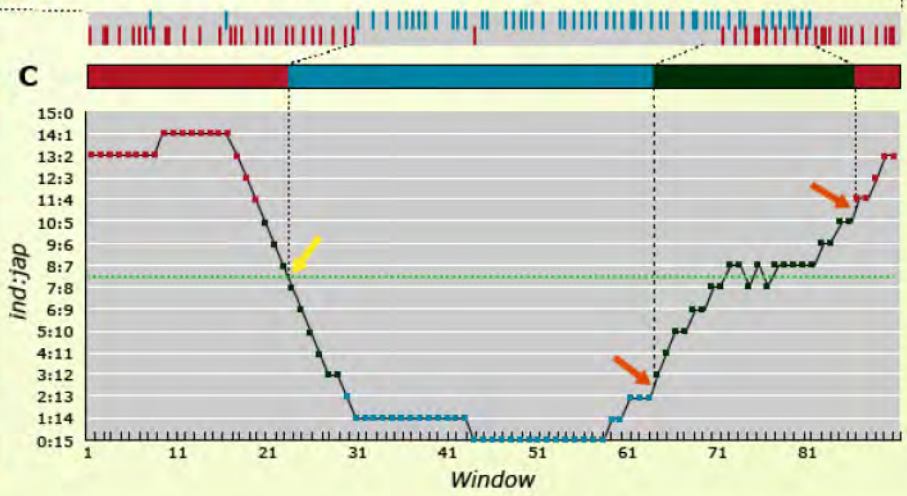
The sliding window (Fig. 10) is a multi-loci mapping algorithm commonly used in association mapping. It involves three steps:
Local haplotypes are inferred from contiguous SNP loci
Genotypes are grouped according to inferred haplotypes
Statistics (F-test) for the genotype-phenotype association and p-values are computed.

Fig. 10 The sliding window approach for data imputation. Adapted from Huang et al., 2010. defined chromosomal regions based on the number of SNPs in a chromosomal region, i.e. defining a window size of w SNPs, and allowing the w to vary according to the size of chromosomal regions showing strong LD. During this process, the window slides along the chromosome at an interval of one SNP until the missing data are inferred for the entire chromosome.



B

ind:jap	Expected probability (%)			Genotype calling
	ind/ind	*ind/jap*	*jap/jap*	
15:0	26.03	0.00	0.00	*ind/ind*
14:1	16.78	0.00	0.00	*ind/ind*
13:2	5.05	0.00	0.00	*ind/ind*
12:3	0.94	0.01	0.00	*ind/ind*
11:4	0.12	0.03	0.00	*ind/ind*
10:5	0.01	0.08	0.00	*ind/jap*
9:6	0.00	0.11	0.00	*ind/jap*
8:7	0.00	0.15	0.00	*ind/jap*
7:8	0.00	0.17	0.00	*ind/jap*
6:9	0.00	0.14	0.00	*ind/jap*
5:10	0.00	0.09	0.00	*ind/jap*
4:11	0.00	0.04	0.00	*ind/jap*
3:12	0.00	0.02	0.01	*ind/jap*
2:13	0.00	0.00	0.24	*jap/jap*
1:14	0.00	0.00	4.84	*jap/jap*
0:15	0.00	0.00	45.15	*jap/jap*
Total	48.94	0.82	50.24	



Data Imputation

THE SLIDING WINDOW APPROACH

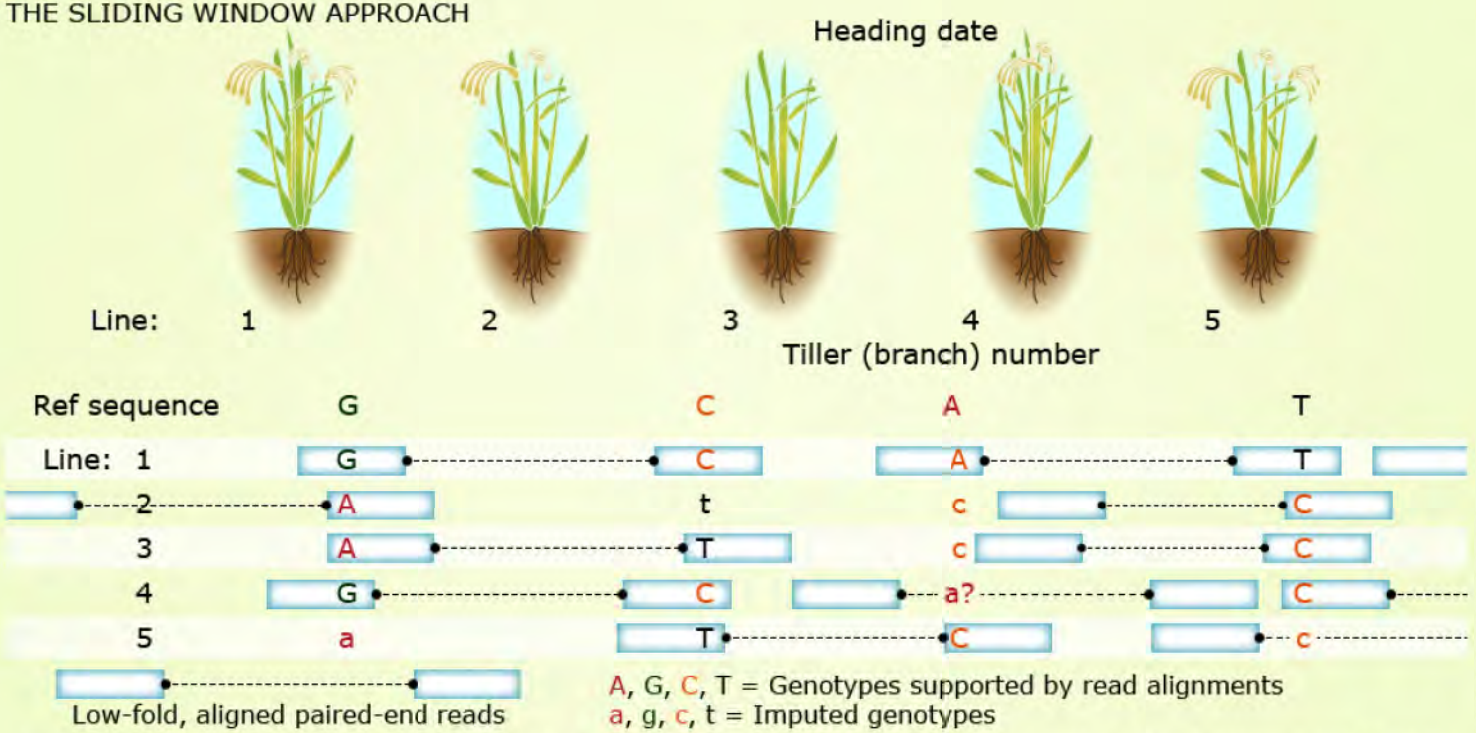


Fig. 11 Construction of a HapMap by NGS sequencing. (a) Variation in agronomic traits (e.g. heading date and tiller branch number) among lines. (b) NGS sequences are aligned to the reference genome for genome-wide genotyping. Aligned reads (gray boxes) facilitate SNP (bases in upper case) identification among lines. Imputation is used to "fill in" missing genotypes (bases in lower case) in areas not covered by sequencing. Boxes with dashed lines are referred to as paired-end reads, and are used to facilitate proper read alignment. Consistent patterns of mismatch between NGS sequence and the reference genome distinguish genetic variability from random sequence variation. Adapted from Clark, 2010.

Data Imputation

BIN MAPS

A bin (Gardiner et al. 1993) is a chromosome segment of about 20 cM flanked by two fixed core markers (a locus or probe that defines a bin boundary). A bin contains all loci within a left fixed core marker to the right fixed core marker. Assigning a locus to a bin is highly dependent on the precision of the mapping data, and increases in likelihood as the number of markers or mapping population increases in size. Bin maps contain coordinates named by the chromosome number, followed by a decimal, and a numeric identifier. 1.00 is the most distal (left or top-most, see arrow in Figure 8) bin on the short arm of the chromosome. At right is the representation of bin boundaries for the 10 maize chromosomes.

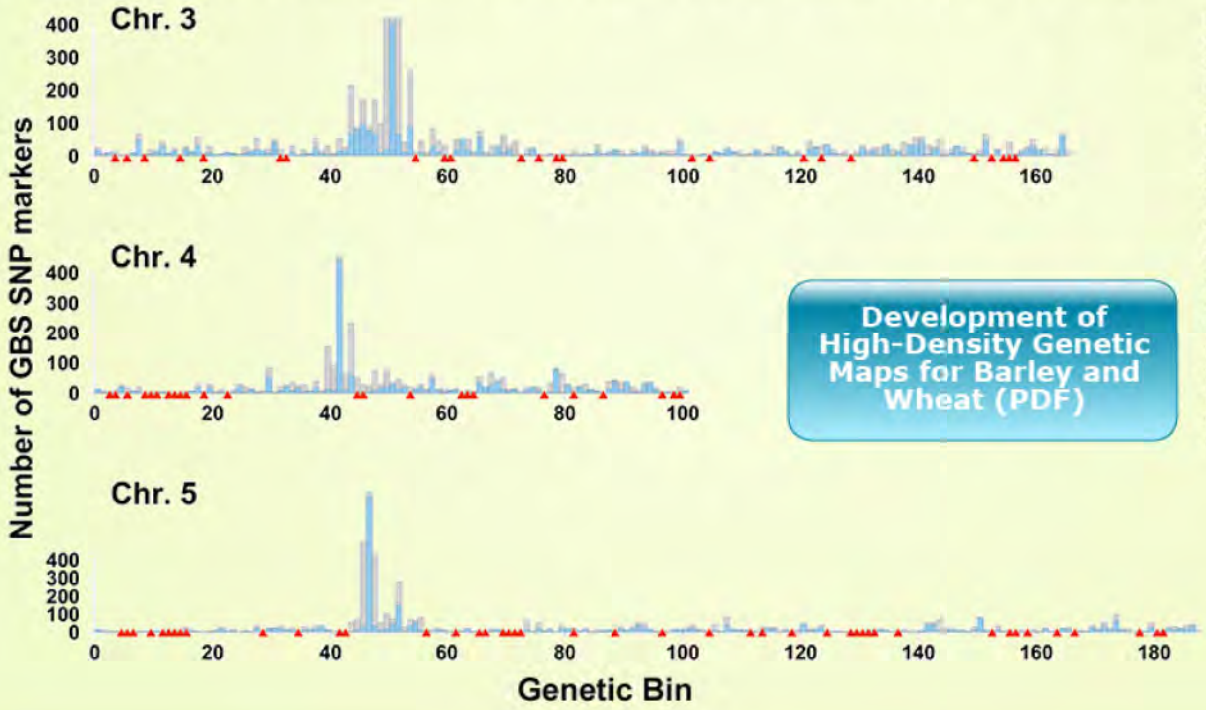
More about bins (Online)

Fig. 12 Bin boundaries for maize chromosomes. The chromosome partitions (white horizontal lines) are based on the concept of bins. Adapted from MaizeGDB.

Data Imputation

BIN MAPS

An example of application of GBS to develop a bin map for a crop with complex genome is provided from work by Poland et al. (2012). In this example, GBS was evaluated in wheat and barley, and a de novo genetic map was constructed using SNP markers from the GBS data (Fig. 13).



Development of High-Density Genetic Maps for Barley and Wheat (PDF)

Fig. 13 Distribution of SNPs discovered by GBS in bin map of barley (only chromosomes 3, 4, and 5 are shown). Histograms represent the number of SNPs from GBS that map to each bin. The number of SNPs mapping to a single bin is represented by the blue bars. SNPs that did not match a particular bin are represented by grey bars. Red triangles below the plots represent bins that failed to match any SNP marker from GBS data. Adapted from Poland et al., 2012.

Data Imputation

TAG SNPs

A tag SNP is a polymorphism in a region of the chromosome with high LD and can be used as a marker for genetic variation without genotyping an entire chromosome.

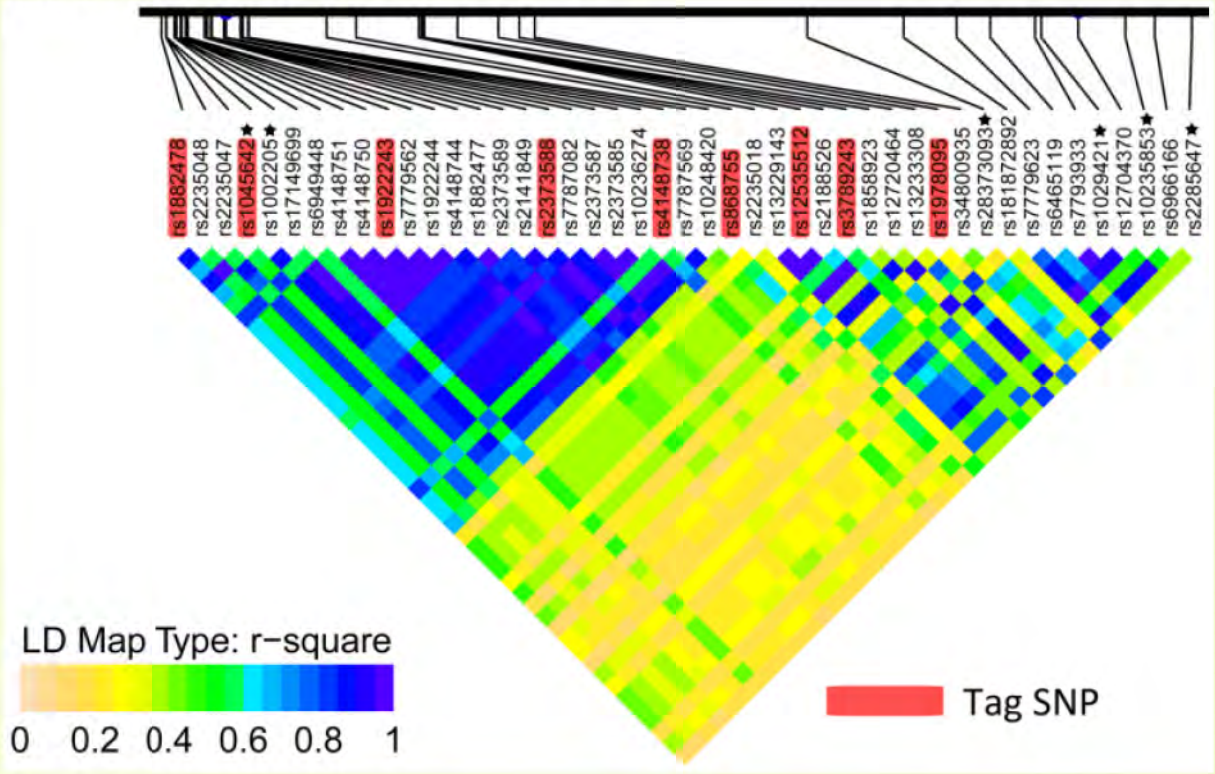


Fig. 14 LD plot of SNPs with top-ranked BFs in CHB of 1000 Genome Phase I. by Weihua Shou,Dazhi Wang, Kaiyue Zhang, Beilan Wang, Zhimin Wang. Licensed under Creative Commons Attribution 3.0 via Wikimedia Commons.

STUDY QUESTIONS

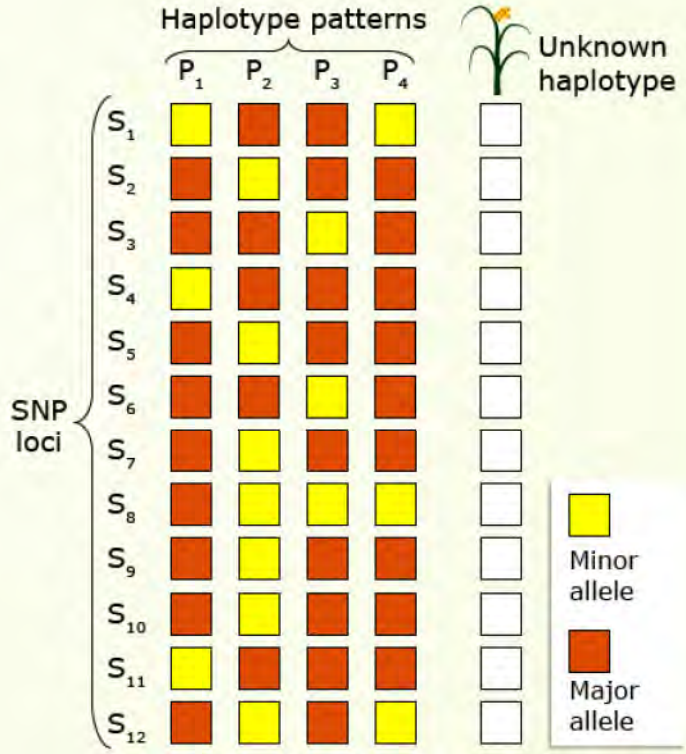
Tag SNPs

In the figure at right, suppose SNPs in two alleles (major and minor) reveal a particular pattern among haplotypes.

What is the minimum number of markers sufficient to discriminate all four haplotypes?

Can the pattern be used to distinguish an unknown haplotype sample?

Yes No



DNA Markers

The surge in the development of new tools for molecular genetics starting in the 1980s made it possible to identify genetic variation at the molecular level based on DNA changes and their impact on the phenotype. Such DNA changes (polymorphisms) can be exploited, e.g., as markers for a particular trait of interest, by plant breeders. Availability of an increasing amount of sequence data from sequencing projects together with new technologies such as next generation sequencing, and bioinformatic tools have reduced the cost of marker discovery and application.

DNA Markers

Molecular or DNA markers reveal sites of variation in DNA. Variability in DNA facilitates the development of markers for mapping and detection or traits. Any DNA sequence can be genetically mapped, like genes leading to plant phenotypes. Prerequisite is, that there must be a polymorphism available for the sequence to be mapped, i.e., two or more different alleles. This can basically be a **single nucleotide polymorphism (SNP)**, a single nucleotide variant at a particular position within the target sequence, or an insertion / deletion (**INDEL**) polymorphism. Target sequences can be amplified by various methods, including **Polymerase chain reaction (PCR)**, and subsequently be visualized to generate "molecular phenotypes" comparable to visual phenotypes, that can be observed by using appropriate equipment. The main use of those SNPs and INDEL polymorphisms is as **molecular markers**. By genetic mapping as described above, linkage between genes affecting agronomic traits or morphological characters, and DNA-based SNP or INDEL markers can be established. It can be more effective in the context of plant breeding, to select indirectly for markers (DNA or non-DNA), than directly for target traits. Reasons can be: lower costs for marker analyses, the ability to run multiple such assays (for DNA markers) in parallel, the ability to select early and to discard undesirable genotypes or to perform selection before flowering, codominant inheritance of markers, among others. Below is a discussion on various types of markers used in plant breeding.

DNA Markers

GENERAL PROPERTIES

DNA markers are readily detectable DNA sequences, whose inheritance can be monitored. The advantage of DNA-based markers is that they are independent of environmental factors. An ideal DNA marker (system) should possess the following properties:

1. Be highly polymorphic
2. Display co-dominant inheritance (to discriminate homozygotes from heterozygotes)
3. Occur at high frequency in the genome
4. Display selective neutral behavior
5. Provide easy access
6. Be simple to evaluate by available set of tools
7. Display high reproducibility, and
8. Facilitate easy exchange of data between laboratories

Historically, DNA markers can be grouped into three main categories: (1) hybridization-based markers, e.g. restriction fragment length polymorphism (RFLP) markers; (2) PCR-based markers, e.g., amplified fragment length polymorphism (AFLP), and simple sequence repeat (SSR); and (3) sequence or chip-based markers, e.g., some procedures for detecting single nucleotide polymorphism (SNP) markers. Examples of molecular markers belonging to the above three categories are further discussed.

Classical DNA Markers

RFLP

Restriction fragment length polymorphism (RFLP) markers involve cutting DNA into fragments and comparing patterns of variability in fragment size, or polymorphisms. RFLP patterns are analyzed by scoring an autoradiograph of a Southern blot. More information about RFLP is found in Crop Genetics eModule8.

Strengths of RFLP

- Co-dominance
- No sequence information is required
- Simplicity not requiring costly instrumentation
- RFLP probe sequences can be used to develop additional markers e.g. Indel
- Transferability across related species

Weaknesses of RFLP

- Analysis requires large amounts of high quality DNA
- Low genotypic throughput (few loci detected per assay)
- Difficult to automate
- Use of radioactive probes restricts the analysis to specific laboratories
- Probes must be physically maintained not allowing sharing between laboratories
- Expensive

Classical DNA Markers

SSR

Simple Sequence Repeat (SSR) markers are widely used markers based upon the high rate of variation in microsatellite loci. SSRs represent a few to hundreds highly variable tandem copies of DNA repeats. Such tandem repeats of usually one to four bases are widespread in higher organisms. Many different microsatellite loci (>100,000) can be present in any plant species. SSRs are a result of slippage during DNA replication or unequal crossover during meiosis.

Variation in SSRs is observed by developing locus-specific primers that anneal to sequences flanking the repeat region; Polymerase Chain Reaction (PCR) is subsequently used to amplify the target region. Alleles (fragments) are visualized as bands with different migration pattern on a gel after electrophoresis. More recently, capillary electrophoresis is used, which also allows to multiplex up to about 16 SSRs per capillary.

The following activity will allow you to use database web tools to search for SSRs and design primers to detect SSR by PCR.

Activity

TRY THIS!

Activity

Develop PCR primers to detect SSRs for maize teosinte branched1.

Step 1: Obtain Zea mays tb1 sequence from **NCBI**.

Step 2: Use the tb1 sequence you obtained from NCBI search for the tb1 locus in MaizeGDB.

Step 3: In your BLAST results view the locations of tb1 in the maize genome.

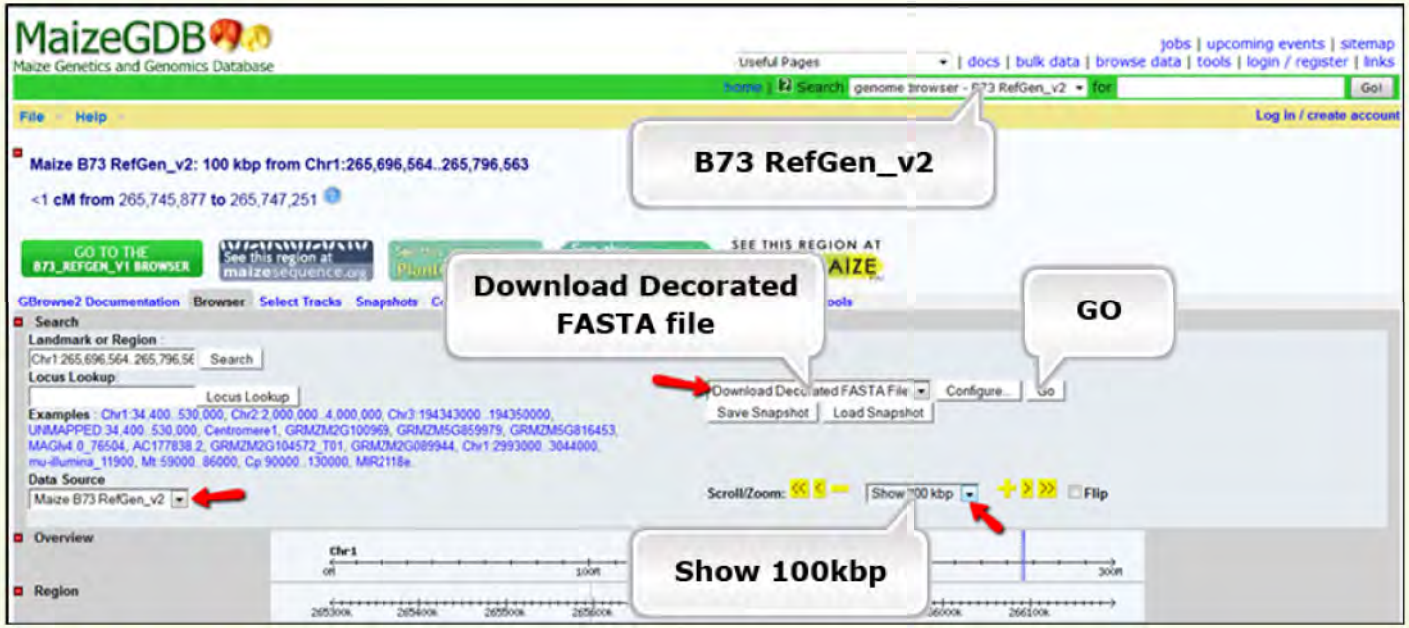
Step 4: Select a hit with an E-value of 0.

Step 4 will take you to this window.

TRY THIS!

Activity

Step 5: Select Maize B73 RefGen_v2, Download Decorated FASTA File and Show 100 kbp. The command to show 100 kbp will allow you to extract about 100 kbp containing the tb1 locus. Last, select GO.



TRY THIS!

Activity

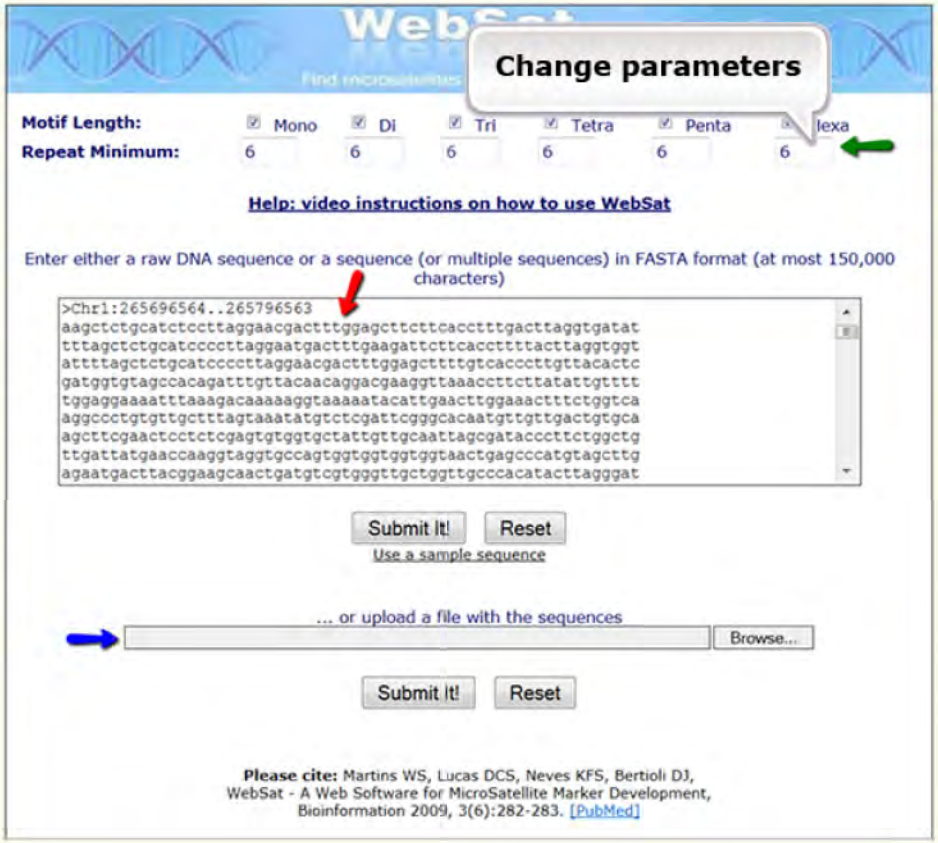
After performing the actions described in Step 5, a new window with the entire 100 kbp sequence you extracted will be displayed. Copy the sequence and proceed to step 6.

```
>Chr1:265696564..265796563
aagctctgcatctccttaggaacgactttggagcttcttcacctttgacttaggtgatat
tttagctctgcatccccttaggaatgactttgaagattcttcacctttttacttaggtggt
attttagctctgcatccccttaggaacgactttggagcttttgtcaccttgttacactc
gatggtgtagccacagatttgttacaacaggacgaaggttaaaccttcttatattgtttt
tggaggaaaatttaaagacaaaaaggtaaaaatacattgaacttggaaactttctggtca
aggccctgtgttgctttagtaaatatgtctcgattcgggcacaatgttgttgactgtgca
agcttcgaactcctctcgagtgtggtgctattgttgcaattagcgatacccttctggctg
ttgattatgaaccaaggtaggtgccagtggtggtggtggtaactgagcccatgtagcttg
agaatgacttacggaagcaactgatgtcgtgggttgctggttgcccacatacttagggat
atatgacaaatatcatgaagcagtatgcaagacttgcttcggctagttctgttgtgcttc
agctttagctatccccttttgcttttggatcgtgacttgacacgtccttgttgtgtgtcc
cttatcctccccacagaataagcagaacaatctcctgagttgagtgctgaaccttccttc
gaagcttcttcctcctctgcctcttggagctggtggcctgtaagaggtttgctacattct
tgaagattgtgagctttgttggctgccttgagtgtgatttgccctgccatcatttttgct
```

TRY THIS!

Activity

Step 6: Several internet-based programs are available for searching microsatellites. For this exercise you will use a web-based software called WebSat. To access WebSat go to wsmartins.net/websat/. You have the option of pasting your copied FASTA sequence or uploading a FASTA file of your sequence. Also, you can change parameters related to the repeat. After entering the 100 kbp tb1 sequence, select "Submit It!"



TRY THIS!

Activity

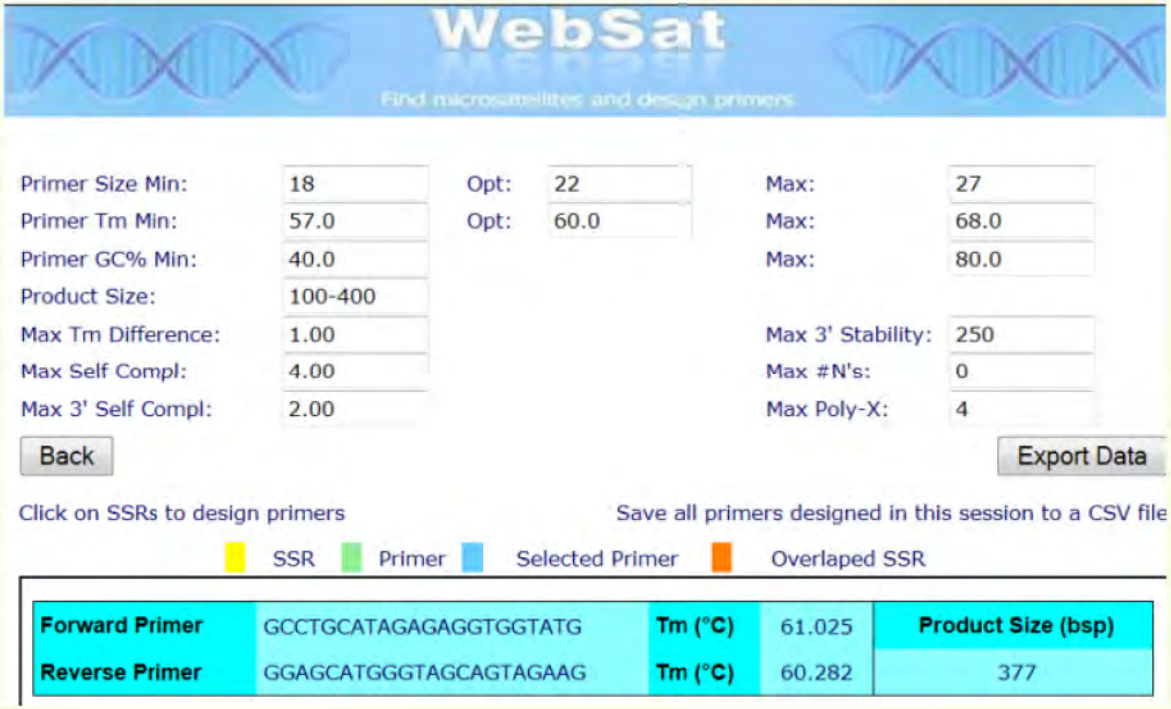
Step 7: After submitting your sequence WebSat will search for SSRs within the sequence and return results with the option to design PCR primers. Design PCR primers around a repeat in the tb1 sequence by selecting a stretch of CTs found between positions 49,141 and 49,351.

```
48791   GTTTCCTGAA GAAGTATTTC ATGGAGGCGC GCACGTCCAT CGTACTGCGT CCTGCAGCTA TGGCCGCCCC
48861   CATCTGGCCA ATAAATGTAC TAGGTCACTT GTAGCCAATA GCGTTTCAAC ATGCACACAG CTTTTCCCCC
48931   AATAGTGCAG GTCCTTGTAT TCTCCTCCCT CTCCCTCACC TCAAATCTCA TCCACACGAA CAGGCGGCAC
49001   GGCAGTATTC CTCCACAGCC CTCCTCTCTA TAAGATGGCA CAGCCCTCTC AGGTAGGGGC GAGTGTCTCA
49071   CTCTCACATA GTAAAAAAAA AAAACGCCCC AAGGTTCTTA AGCACAATTC TCTAGCTATC TTGGTCTCCT
49141   ACACAGCCTA TGCACATGAG CCCATGCCTC TCCTCTCCTT GCGCCTGCAT AGAGAGGTGG TATGATCACC
49211   TGGAAAGTTT TTAACTCTCT CTCTCTCTCT CTCTCTCTCT CTCTCTCTTA CAAGCCTAGA CCTTATGCAT
49281   GGTCGGACGG ACACATCTGA TCATAGGACA TATGAGTAGG CCACACTCCT CCTGCCCCTC TCTCGTAGAG
49351   ATCAACACAC ACTGCTCTTA GTGCCAGGAC CTAGAGAGGG GAGCGTGGAG AGGGCATCAG GGGGCCTTGG
49421   AGTCCCATCA GTAAAGCACA TGTTTCCTTT CTGTGATTCC TCAAGCCCCA TGGACTTACC GCTTTACCAA
49491   CAACTGCAGC TAAGCCCGTC TTCCCCAAAG ACGGACCAAT CCAGCAGCTT CTACTGCTAC CCATGCTCCC
49561   CTCCCTTCGC CGCCGCCGAC GCCAGCTTTC CCCTCAGCTA CCAGATCGGT AGTGCCGCGG CCGCCGACGC
49631   CACCCCTCCA CAAGCCGTGA TCAACTCGCC GGACCTGCCG GTGCAGGCGC TGATGGACCA CGCGCCGGCG
```

TRY THIS!

Activity

WebSat selects primers that fit the parameters of your choice to flank the SSR of your interest. In this case the forward and reverse primers around a stretch of CTs you selected will result in a PCR fragment of about 377 bp.



Classical DNA Markers

SSR

Strength of SSR markers:

- Hypervariable, multiple alleles (high PIC)
- In silico development straightforward

Weakness of SSR markers:

- Capability for multiplexing limited (max. 10-15)
- Affects costs/datapoint
- Few intragenic SSRs

More information online:

Transferability of molecular markers can help increase resolution of genomes that are not well characterized

SSR markers located within genes can be used for direct selection of an allele

Classical DNA Markers

AFLP

Amplified Fragment Length Polymorphism (AFLP) markers combine RFLP and PCR. In AFLP genomic DNA is digested with restriction enzymes followed a ligation step where adapters are added to both ends of the restriction fragments. PCR is carried out on the adapter-ligated mixture, using primers that target the adapter, but that vary in the base(s) at the 3' end of the primer. Figure 15 in the text (see link below) describes the steps to detect and analyze AFLP markers.

Strength of AFLP markers:

- High marker index
- Amenable for automation
- Robust
- No prior sequence information required
- Special applications: Gene family profiling; Methylation assay
- Established service company: KeyGene

Weakness of AFLP markers:

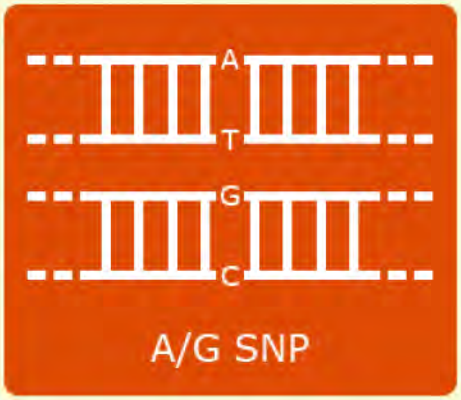
- Random loci, might differ between populations
- Dominant marker system

Another example of AFLP (Online)

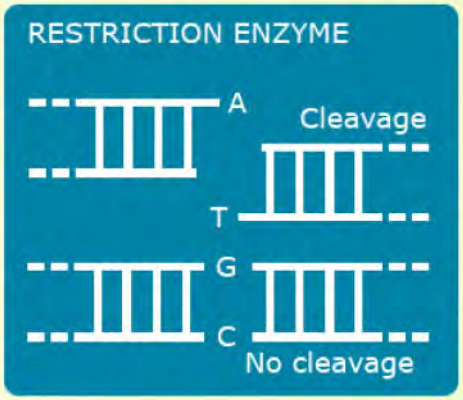
Current DNA Markers

SNP

Single nucleotide polymorphisms (SNPs) are the most abundant kind of polymorphisms in eukaryotic genomes. SNPs are single nucleotide differences (transition or transversion) between allelic sequences. SNPs might cause polymorphisms detectable as RFLP or AFLP markers, if they occur in restriction enzyme recognition sites. Some principles exploited in SNP detection are shown in Figure 15.



In Figure 15A a A to G transition is described, including the various methods that can be used to detect the A- and the G-alleles.



In Figure 15B, restriction enzymes are used to detect the polymorphism of the target DNA. The probe that perfectly matches the target will be stable, and a mismatch would be unstable. The probes are usually labeled with a fluorescing dye or radioisotope for the detection by a laser scanner, or auto-radiography (e.g., Southern blot analysis).

Current DNA Markers

SNP

Overall, many different genotyping approaches are available ranging from low to high throughput. Some platforms permit users to pick custom SNPs but the highest throughput assays are available only in fixed contents. Not all custom SNPs will work for every format and multiple SNPs may be required to carry out most projects targeting specific SNPs. However, there are still trade-offs for throughput, that is, samples versus SNPs to be analyzed. Ultimately, cost will dictate how a SNP project is designed. Regardless of the study, design, quality control and tracking are critical to the success of the project. Laboratory Information Management Systems (LIMS) are important in every study design. The following are examples of SNP genotyping systems that are commonly used by plant breeders.

Current DNA Markers

SNP: TAQMAN ASSAY

TaqMan SNP assays (Figure 16) are based on PCR using four oligonucleotide primers: (1) A set of forward and reverse primers that are designed and tested for each SNP, and (2) Two hydrolysis (Taqman) assay probes conjugated with fluorescent dyes and quenchers. Taqman probes are designed to anneal within a region of the PCR fragment resulting from the forward and reverse primers. The quencher ensures that a dye does not fluoresce before Taqman probes have annealed to their target during PCR. The PCR reaction is catalyzed by a polymerase enzyme with 5′ to 3′ exonuclease activity. The 5′ to 3′ exonuclease activity is required to cleave the quencher from the dye allowing fluorescence to be produced during PCR amplification.

An A to G transition is shown within the target DNA (DNA template).

Two Taqman probes are designed to recognize either A or G. The probes are linked with dyes and quenchers.

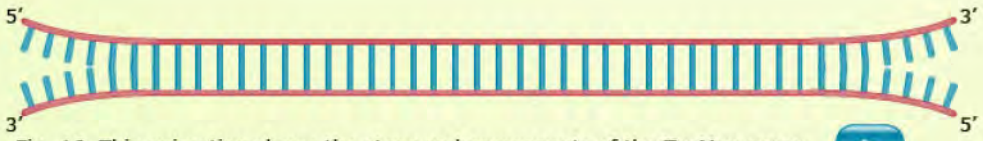


Fig. 16 This animation shows the steps and components of the TaqMan assay.

Current DNA Markers

SNP: SEQUENOM MASSARRAY SYSTEM

The Sequenom MassArray system (Fig. 17) uses highly multiplexed PCR reactions to screen multiple mutation sites simultaneously by primer extension combined with Matrix-Assisted Laser Desorption/Ionization-Time of Flight mass spectrometry (MALDI-TOF-MS). The system provides rapid and quantitative readout allowing detection of mutations, gene copy number, methylation status, and level of expression of allelic variants. Up to about 20 SNPs multiplied by about 400 samples can be analyzed at a time. The current costs are about $0.10 per datapoint.

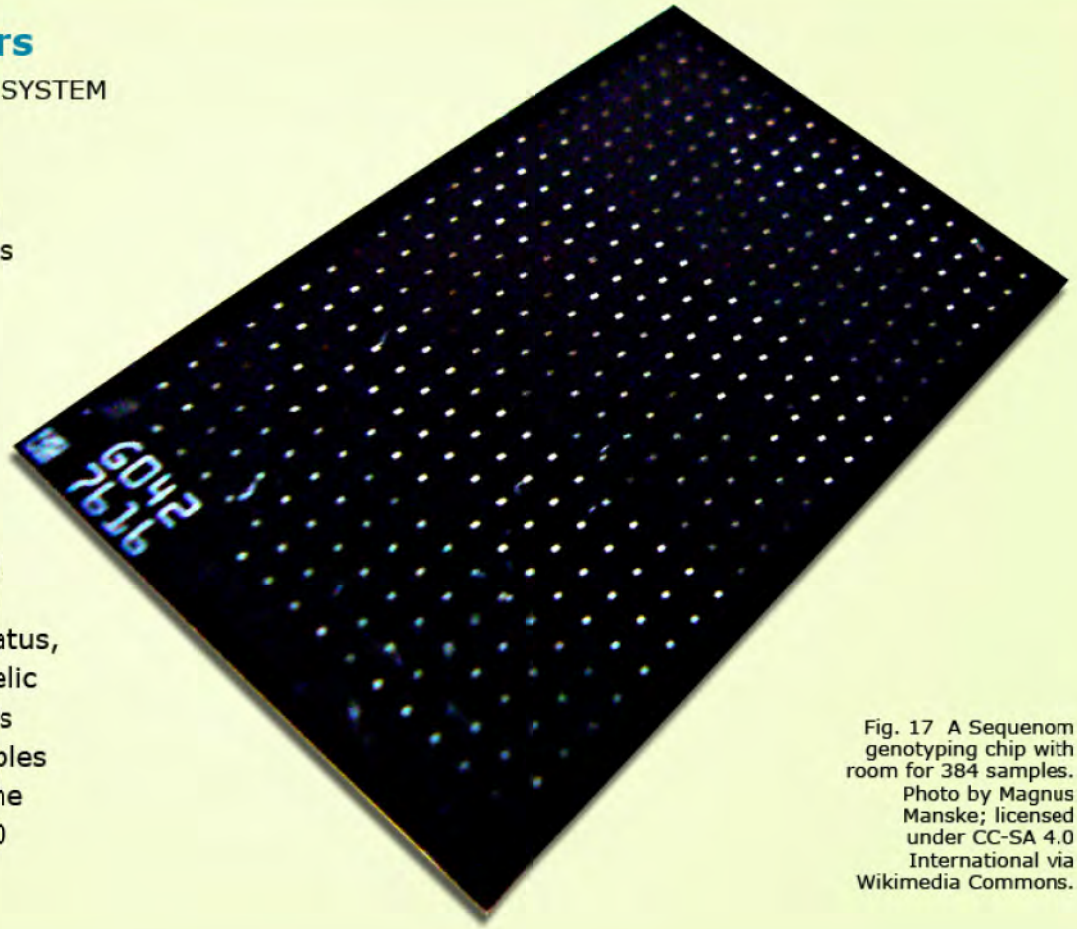


Fig. 17 A Sequenom genotyping chip with room for 384 samples. Photo by Magnus Manske; licensed under CC-SA 4.0 International via Wikimedia Commons.

Current DNA Markers

SNP: SEQUENOM MASSARRAY SYSTEM

The three key steps in SNP analysis using the Sequenom system are (1) Target amplification (2) Primer extension and (3) Signal detection and ratio analysis. These steps are further described in Fig. 18.

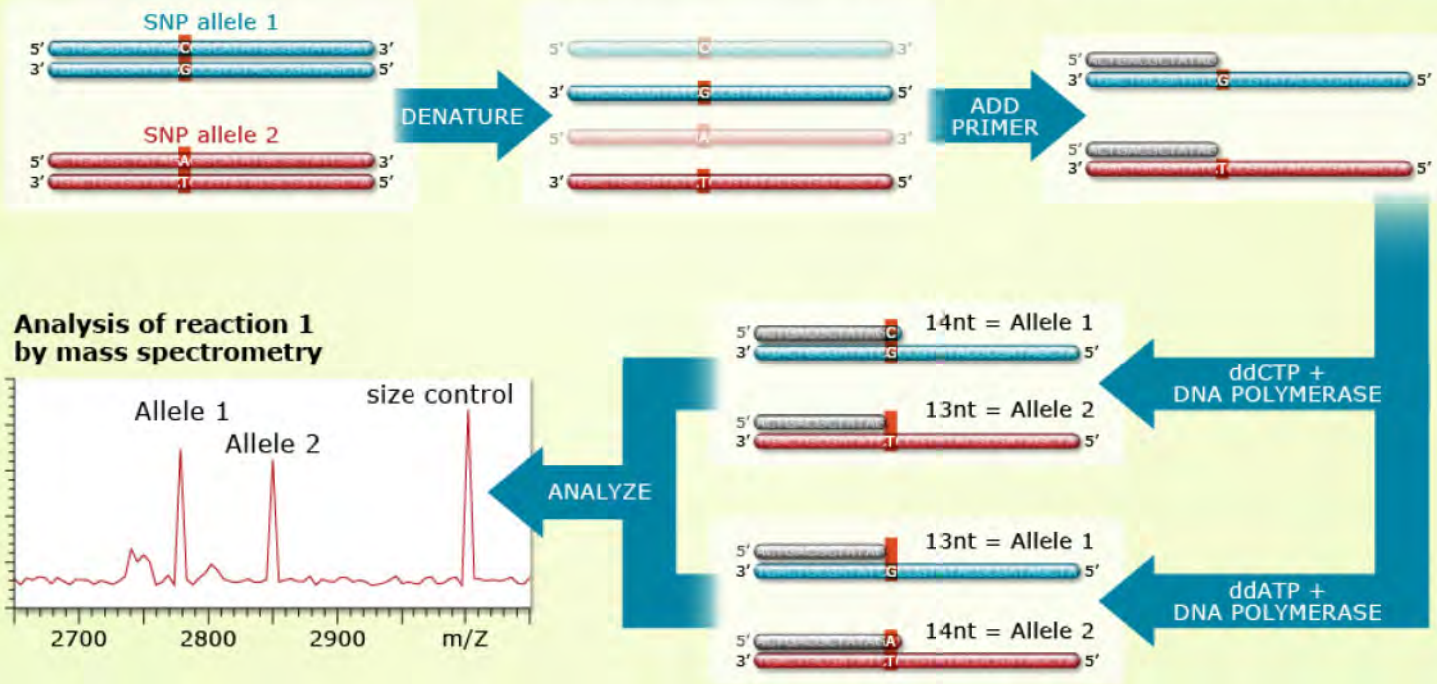


Fig. 18 SNP detection with DNA polymerase-assisted single nucleotide primer extension.

Current DNA Markers

SNP: GOLDENGATE ASSAY

The GoldenGate assay (Fig. 19) involves the addition of biotin to genomic DNA to immobilize the DNA on avidin-coated particles which bind biotin. The assay uses three oligonucleotide primers, two of these (P_1 and P_2) are specific for the two SNP alleles, and the third (P_3) is a locus-specific primer that is tagged with a sequence for capture on solid support. In the reaction, the allele- and locus-specific primers anneal with the genomic DNA followed by extension using a DNA polymerase. After extension, the products are ligated to the tag sequence by a ligase. PCR primers containing fluorescence labels recognize the P_1, P_2, P_3 sequences. The extension products containing the fluorescence labels are captured on the BeadArray containing complementary tag sequences for fluorescence detection.

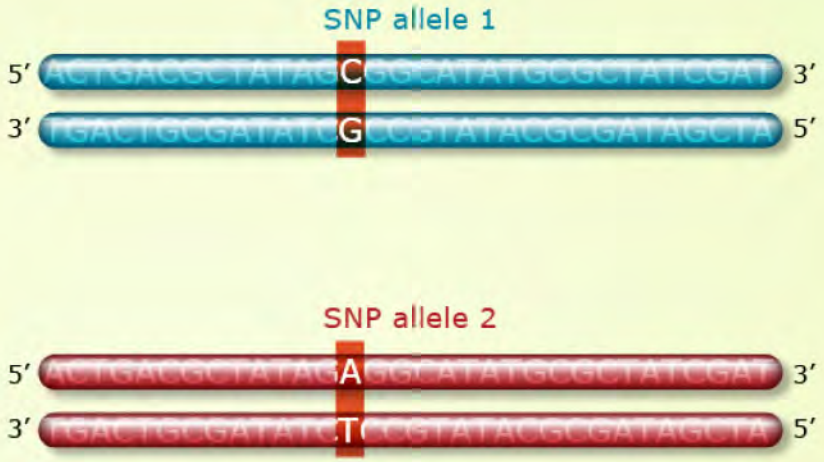


Fig. 19 The reaction principle and steps of the GoldenGate assay. Adapted from Illumina (2006)

Current DNA Markers

SNP: GOLDENGATE ASSAY

The GoldenGate assay (Fig. 19) involves the addition of biotin to genomic DNA to immobilize the DNA on avidin-coated particles which bind biotin. The assay uses three oligonucleotide primers, two of these (P_1 and P_2) are specific for the two SNP alleles, and the third (P_3) is a locus-specific primer that is tagged with a sequence for capture on solid support. In the reaction, the allele- and locus-specific primers anneal with the genomic DNA followed by extension using a DNA polymerase. After extension, the products are ligated to the tag sequence by a ligase. PCR primers containing fluorescence labels recognize the P_1, P_2, P_3 sequences. The extension products containing the fluorescence labels are captured on the BeadArray containing complementary tag sequences for fluorescence detection.

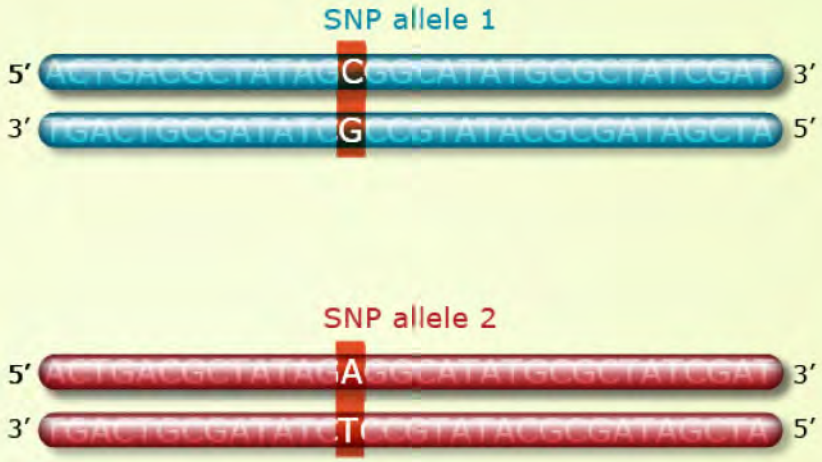


Fig. 19 The reaction principle and steps of the GoldenGate assay. Adapted from Illumina (2006)

Current DNA Markers

SNP: GOLDENGATE ASSAY

The GoldenGate assay (Fig. 19) involves the addition of biotin to genomic DNA to immobilize the DNA on avidin-coated particles which bind biotin. The assay uses three oligonucleotide primers, two of these (P_1 and P_2) are specific for the two SNP alleles, and the third (P_3) is a locus-specific primer that is tagged with a sequence for capture on solid support. In the reaction, the allele- and locus-specific primers anneal with the genomic DNA followed by extension using a DNA polymerase. After extension, the products are ligated to the tag sequence by a ligase. PCR primers containing fluorescence labels recognize the P_1, P_2, P_3 sequences. The extension products containing the fluorescence labels are captured on the BeadArray containing complementary tag sequences for fluorescence detection.

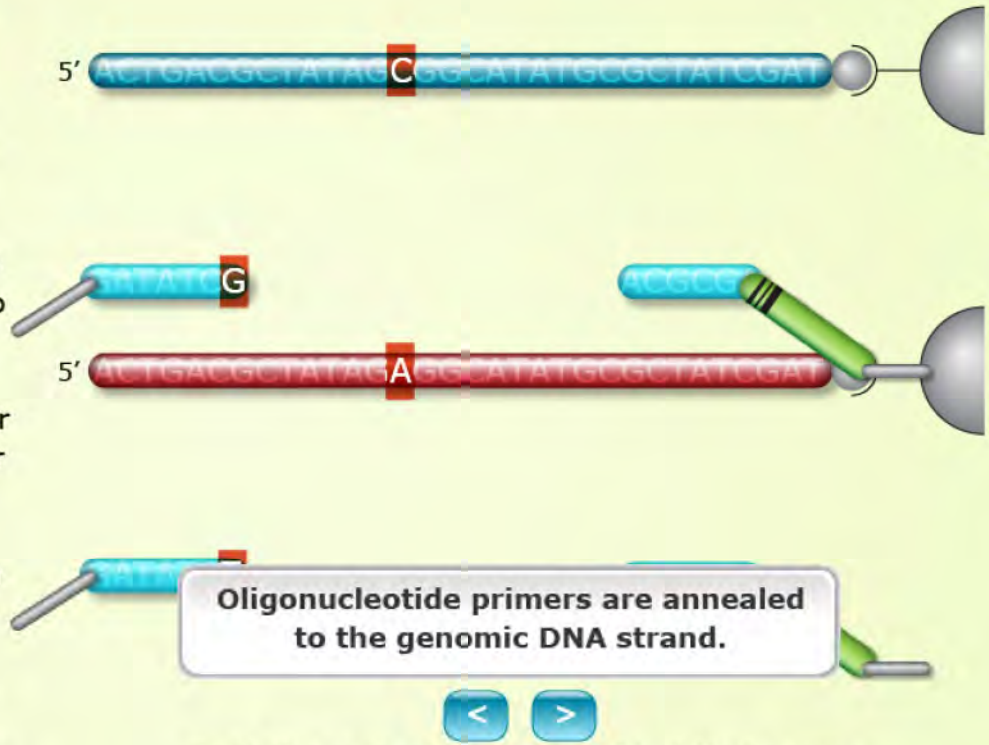


Fig. 19 The reaction principle and steps of the GoldenGate assay. Adapted from Illumina (2006)

Current DNA Markers

SNP: GOLDENGATE ASSAY

The GoldenGate assay (Fig. 19) involves the addition of biotin to genomic DNA to immobilize the DNA on avidin-coated particles which bind biotin. The assay uses three oligonucleotide primers, two of these (P_1 and P_2) are specific for the two SNP alleles, and the third (P_3) is a locus-specific primer that is tagged with a sequence for capture on solid support. In the reaction, the allele- and locus-specific primers anneal with the genomic DNA followed by extension using a DNA polymerase. After extension, the products are ligated to the tag sequence by a ligase. PCR primers containing fluorescence labels recognize the P_1, P_2, P_3 sequences. The extension products containing the fluorescence labels are captured on the BeadArray containing complementary tag sequences for fluorescence detection.

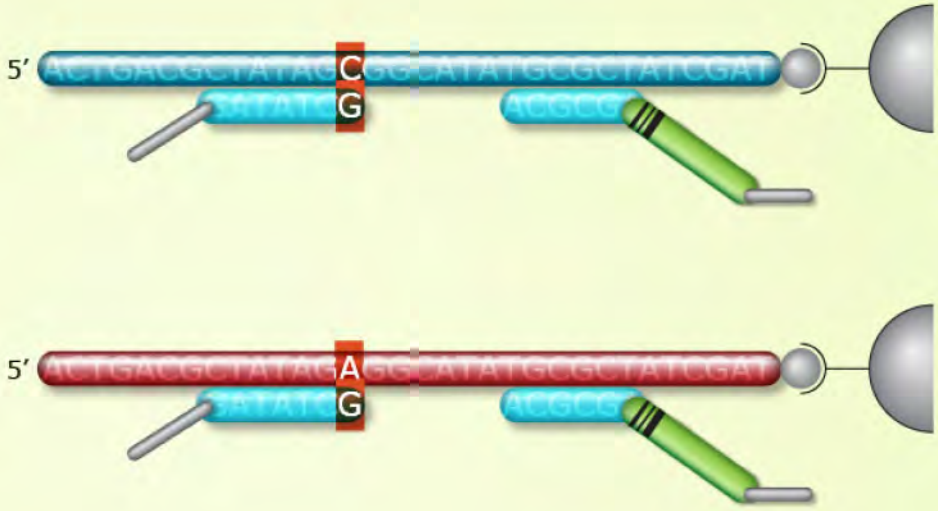


Fig. 19 The reaction principle and steps of the GoldenGate assay. Adapted from Illumina (2006)

Current DNA Markers

SNP: GOLDENGATE ASSAY

The GoldenGate assay (Fig. 19) involves the addition of biotin to genomic DNA to immobilize the DNA on avidin-coated particles which bind biotin. The assay uses three oligonucleotide primers, two of these (P_1 and P_2) are specific for the two SNP alleles, and the third (P_3) is a locus-specific primer that is tagged with a sequence for capture on solid support. In the reaction, the allele- and locus-specific primers anneal with the genomic DNA followed by extension using a DNA polymerase. After extension, the products are ligated to the tag sequence by a ligase. PCR primers containing fluorescence labels recognize the P_1, P_2, P_3 sequences. The extension products containing the fluorescence labels are captured on the BeadArray containing complementary tag sequences for fluorescence detection.

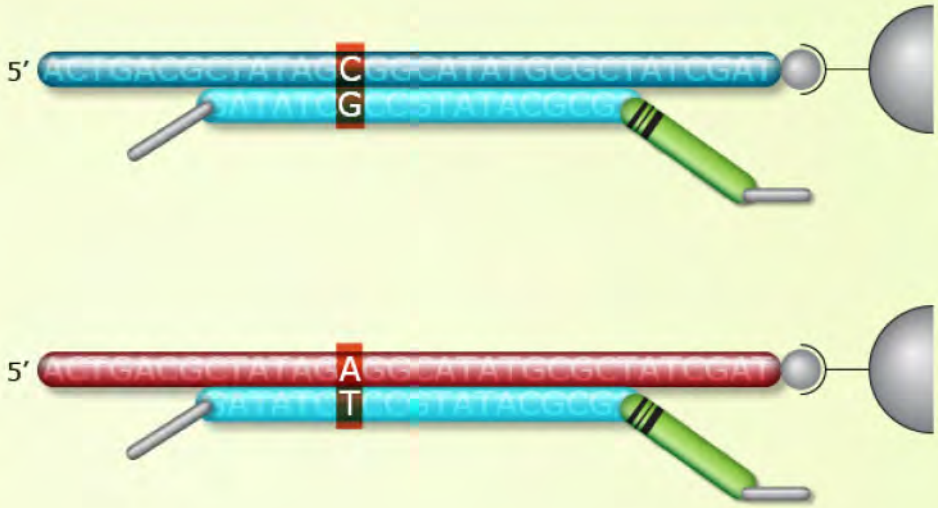


Fig. 19 The reaction principle and steps of the GoldenGate assay. Adapted from Illumina (2006)

Current DNA Markers

SNP: GOLDENGATE ASSAY

The GoldenGate assay (Fig. 19) involves the addition of biotin to genomic DNA to immobilize the DNA on avidin-coated particles which bind biotin. The assay uses three oligonucleotide primers, two of these (P_1 and P_2) are specific for the two SNP alleles, and the third (P_3) is a locus-specific primer that is tagged with a sequence for capture on solid support. In the reaction, the allele- and locus-specific primers anneal with the genomic DNA followed by extension using a DNA polymerase. After extension, the products are ligated to the tag sequence by a ligase. PCR primers containing fluorescence labels recognize the P_1, P_2, P_3 sequences. The extension products containing the fluorescence labels are captured on the BeadArray containing complementary tag sequences for fluorescence detection.

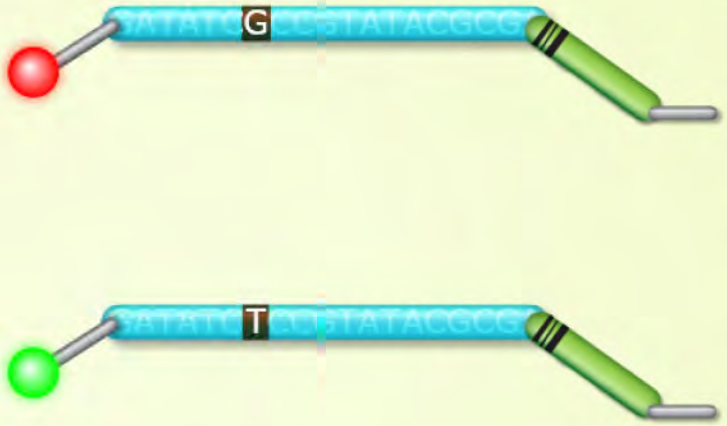


Fig. 19 The reaction principle and steps of the GoldenGate assay. Adapted from Illumina (2006)

Current DNA Markers

SNP: GOLDENGATE ASSAY

The GoldenGate assay (Fig. 19) involves the addition of biotin to genomic DNA to immobilize the DNA on avidin-coated particles which bind biotin. The assay uses three oligonucleotide primers, two of these (P_1 and P_2) are specific for the two SNP alleles, and the third (P_3) is a locus-specific primer that is tagged with a sequence for capture on solid support. In the reaction, the allele- and locus-specific primers anneal with the genomic DNA followed by extension using a DNA polymerase. After extension, the products are ligated to the tag sequence by a ligase. PCR primers containing fluorescence labels recognize the P_1, P_2, P_3 sequences. The extension products containing the fluorescence labels are captured on the BeadArray containing complementary tag sequences for fluorescence detection.

Fig. 19 The reaction principle and steps of the GoldenGate assay. Adapted from Illumina (2006)

Current DNA Markers

SNP: GOLDENGATE ASSAY

The GoldenGate assay (Fig. 19) involves the addition of biotin to genomic DNA to immobilize the DNA on avidin-coated particles which bind biotin. The assay uses three oligonucleotide primers, two of these (P_1 and P_2) are specific for the two SNP alleles, and the third (P_3) is a locus-specific primer that is tagged with a sequence for capture on solid support. In the reaction, the allele- and locus-specific primers anneal with the genomic DNA followed by extension using a DNA polymerase. After extension, the products are ligated to the tag sequence by a ligase. PCR primers containing fluorescence labels recognize the P_1, P_2, P_3 sequences. The extension products containing the fluorescence labels are captured on the BeadArray containing complementary tag sequences for fluorescence detection.

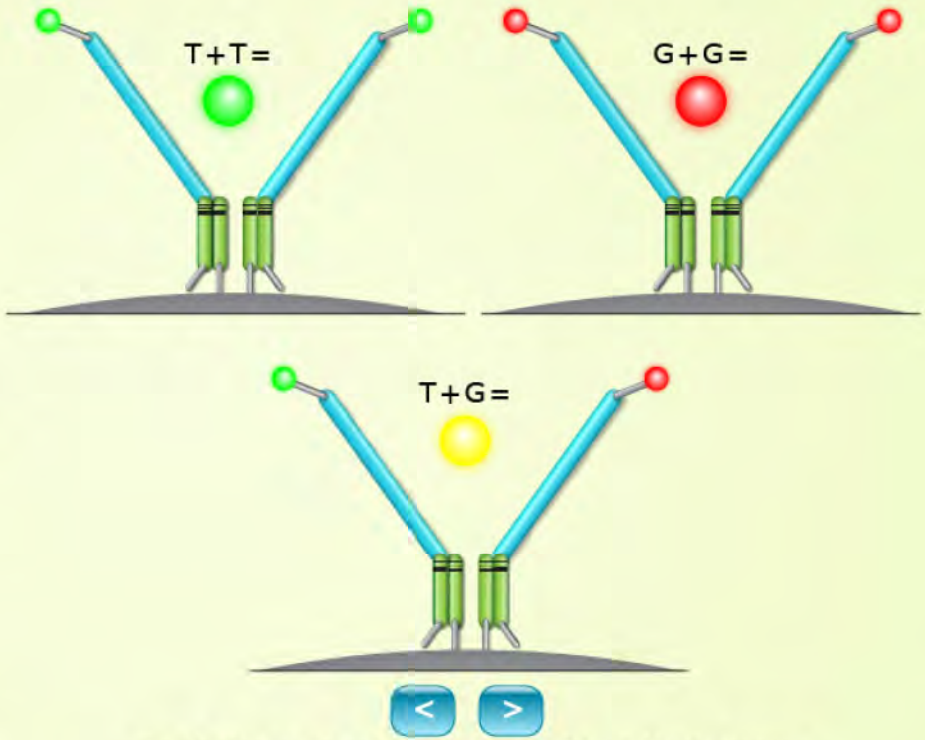


Fig. 19 The reaction principle and steps of the GoldenGate assay. Adapted from Illumina (2006)

Current DNA Markers

SNP: GOLDENGATE ASSAY

The GoldenGate assay (Fig. 19) involves the addition of biotin to genomic DNA to immobilize the DNA on avidin-coated particles which bind biotin. The assay uses three oligonucleotide primers, two of these (P_1 and P_2) are specific for the two SNP alleles, and the third (P_3) is a locus-specific primer that is tagged with a sequence for capture on solid support. In the reaction, the allele- and locus-specific primers anneal with the genomic DNA followed by extension using a DNA polymerase. After extension, the products are ligated to the tag sequence by a ligase. PCR primers containing fluorescence labels recognize the P_1, P_2, P_3 sequences. The extension products containing the fluorescence labels are captured on the BeadArray containing complementary tag sequences for fluorescence detection.

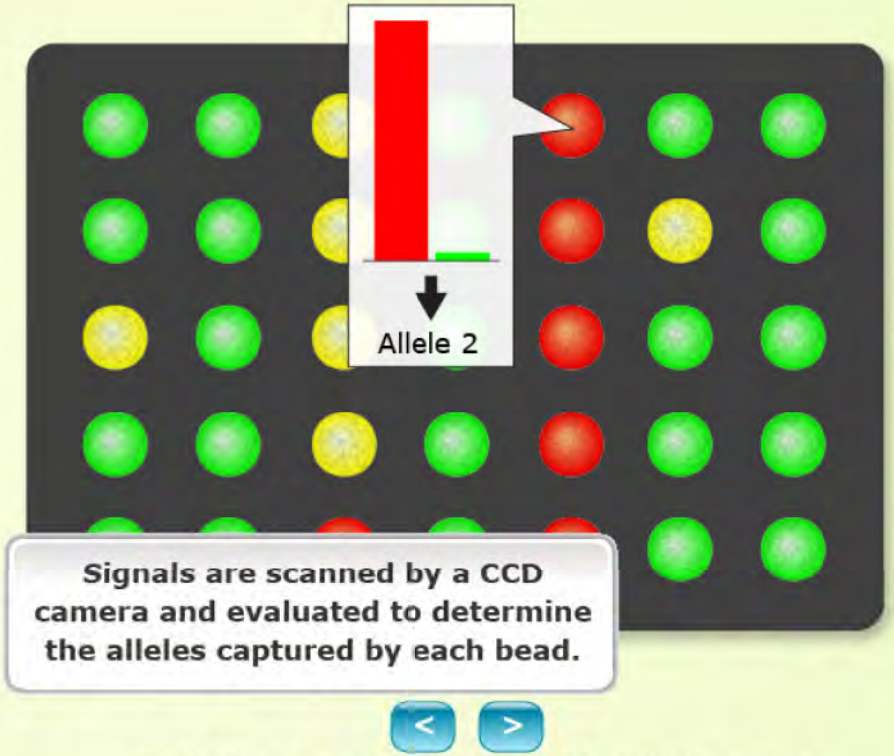


Fig. 19 The reaction principle and steps of the GoldenGate assay. Adapted from Illumina (2006)

Current DNA Markers

SNP: GOLDENGATE ASSAY

The GoldenGate assay (Fig. 19) involves the addition of biotin to genomic DNA to immobilize the DNA on avidin-coated particles which bind biotin. The assay uses three oligonucleotide primers, two of these (P_1 and P_2) are specific for the two SNP alleles, and the third (P_3) is a locus-specific primer that is tagged with a sequence for capture on solid support. In the reaction, the allele- and locus-specific primers anneal with the genomic DNA followed by extension using a DNA polymerase. After extension, the products are ligated to the tag sequence by a ligase. PCR primers containing fluorescence labels recognize the P_1, P_2, P_3 sequences. The extension products containing the fluorescence labels are captured on the BeadArray containing complementary tag sequences for fluorescence detection.

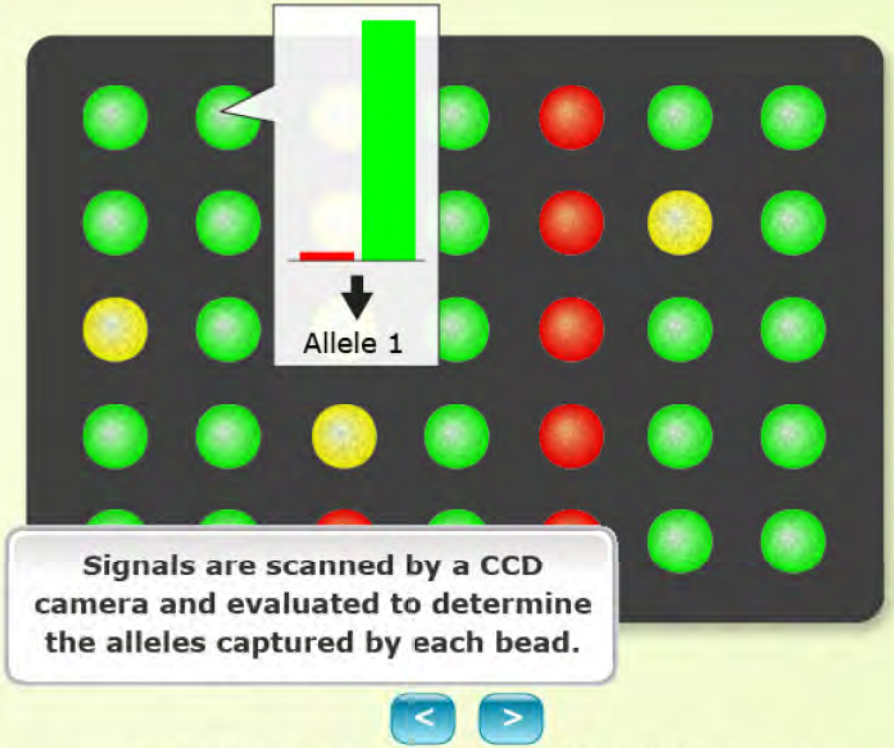


Fig. 19 The reaction principle and steps of the GoldenGate assay. Adapted from Illumina (2006)

Current DNA Markers

SNP: GOLDENGATE ASSAY

The GoldenGate assay (Fig. 19) involves the addition of biotin to genomic DNA to immobilize the DNA on avidin-coated particles which bind biotin. The assay uses three oligonucleotide primers, two of these (P_1 and P_2) are specific for the two SNP alleles, and the third (P_3) is a locus-specific primer that is tagged with a sequence for capture on solid support. In the reaction, the allele- and locus-specific primers anneal with the genomic DNA followed by extension using a DNA polymerase. After extension, the products are ligated to the tag sequence by a ligase. PCR primers containing fluorescence labels recognize the P_1, P_2, P_3 sequences. The extension products containing the fluorescence labels are captured on the BeadArray containing complementary tag sequences for fluorescence detection.

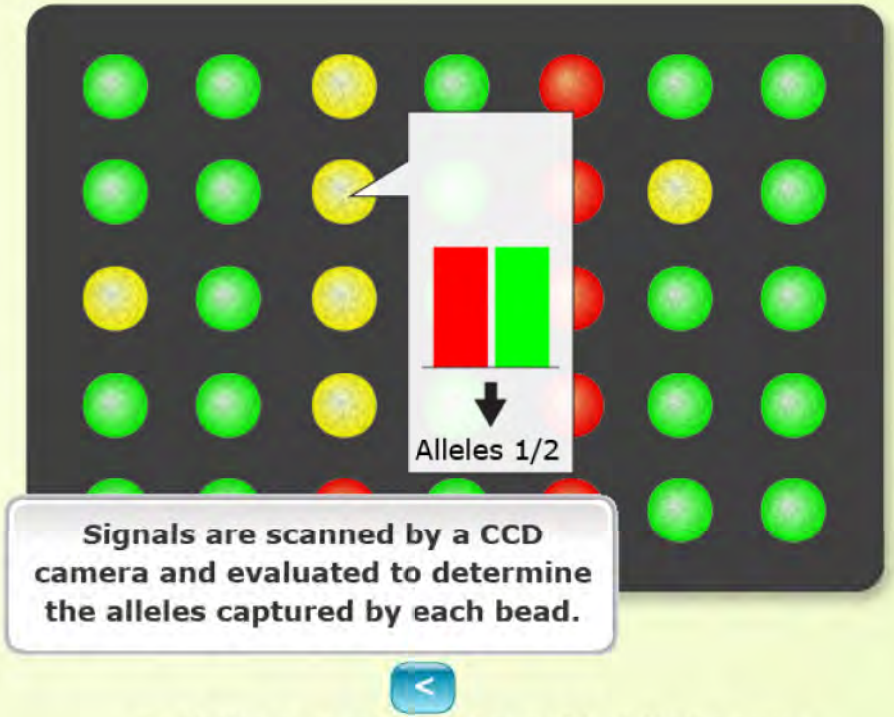


Fig. 19 The reaction principle and steps of the GoldenGate assay. Adapted from Illumina (2006)

Current DNA Markers

INDEL

Insertions and deletions (Indels) cause changes in DNA sequence by deletion or insertion. Indels can range in size from one or few bases to multiple megabases. Small deletions from a few base pairs to kilobases in length most often arise from unequal crossover during meiosis.

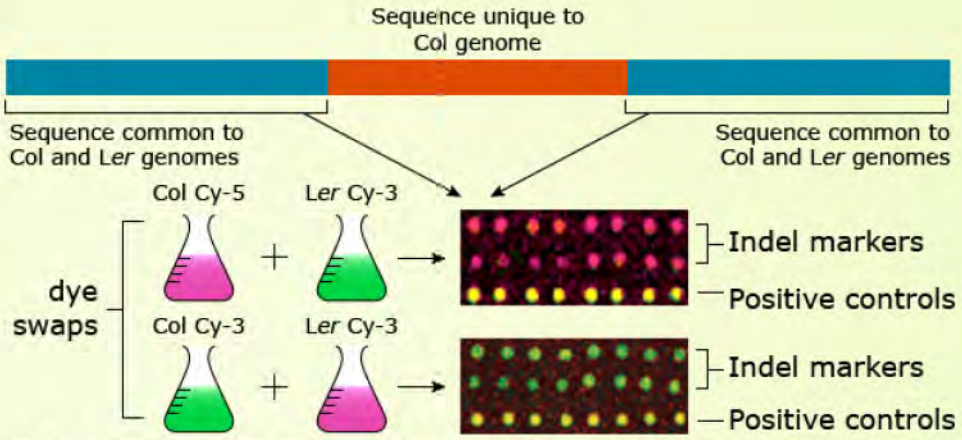


Fig. 20 Arabidopsis INDEL array technology. The array surface is coated with 70 bp indel oligonucleotides unique to the Columbia ecotype. The DNA from Col and Ler is labeled with fluorescent dyes. The labeled DNA is hybridized to the array. Adapted from Salathia et al., 2007.

The Arabidopsis INDEL array (Salathia et al. 2007) is a microarray-based system (Fig. 20) that can be used to assess up to 240 polymorphic markers by hybridization. The array is based on 70-mer oligonucleotides of indels present in two Arabidopsis ecotypes, Columbia-0 (Col-0) and Landsberg erecta (Ler). PCR primers are also available for validation of array-based data. Groups of 16 lines can be genotyped together in a single experiment.

FURTHER THOUGHT

Activity

As show in Figure 20, Salathia et al. (2007) swapped the dye labels for Col and Ler. What is the significance of dye swaps?

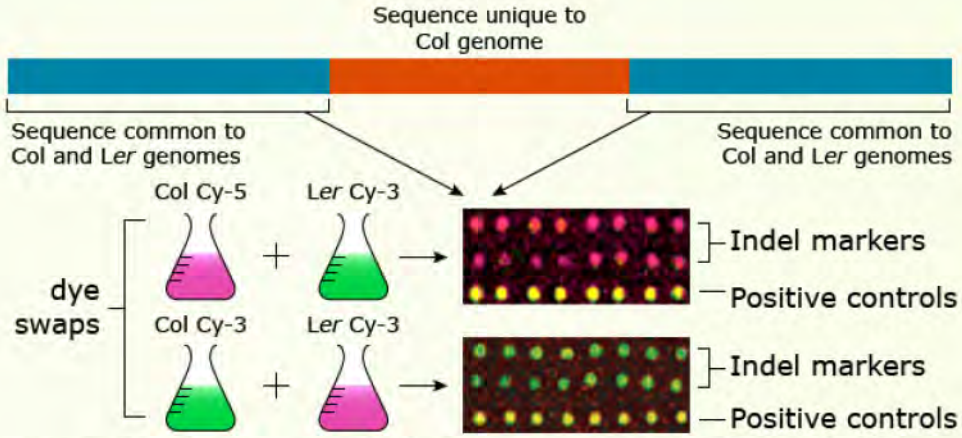


Fig. 20 Arabidopsis INDEL array technology. The array surface is coated with 70 bp indel oligonucleotides unique to the Columbia ecotype. The DNA from Col and *Ler* is labeled with fluorescent dyes. The labeled DNA is hybridized to the array. Adapted from Salathia et al., 2007.

General Application of Markers

GENETIC FINGERPRINTING

Genetic fingerprinting is a method that employs the uniqueness of DNA to classify individuals in distinct or similar groups. Based on the fact that genomes of different individuals will contain polymorphisms, a particular DNA profile can be established for a particular organism. This profile is specific to that individual, and as unique as a fingerprint.

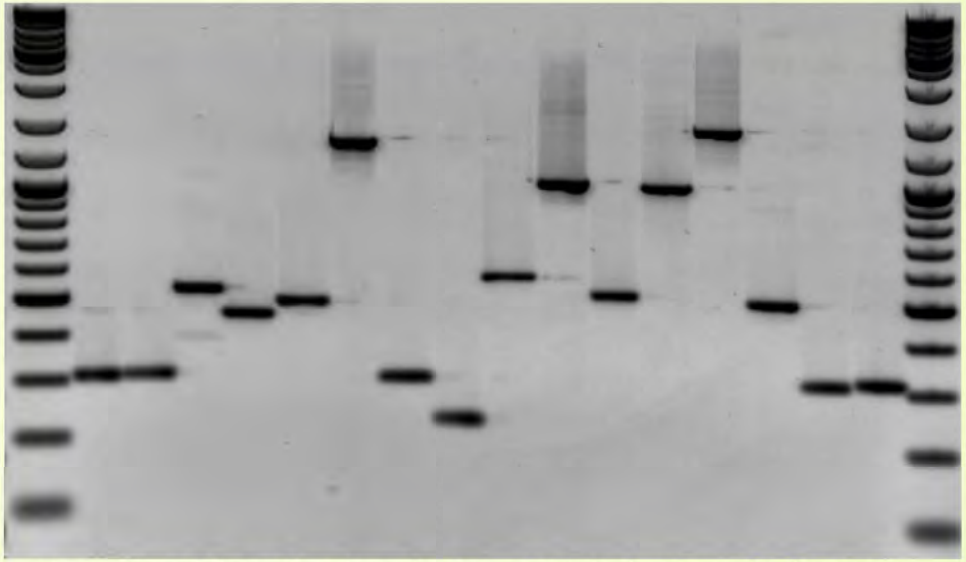
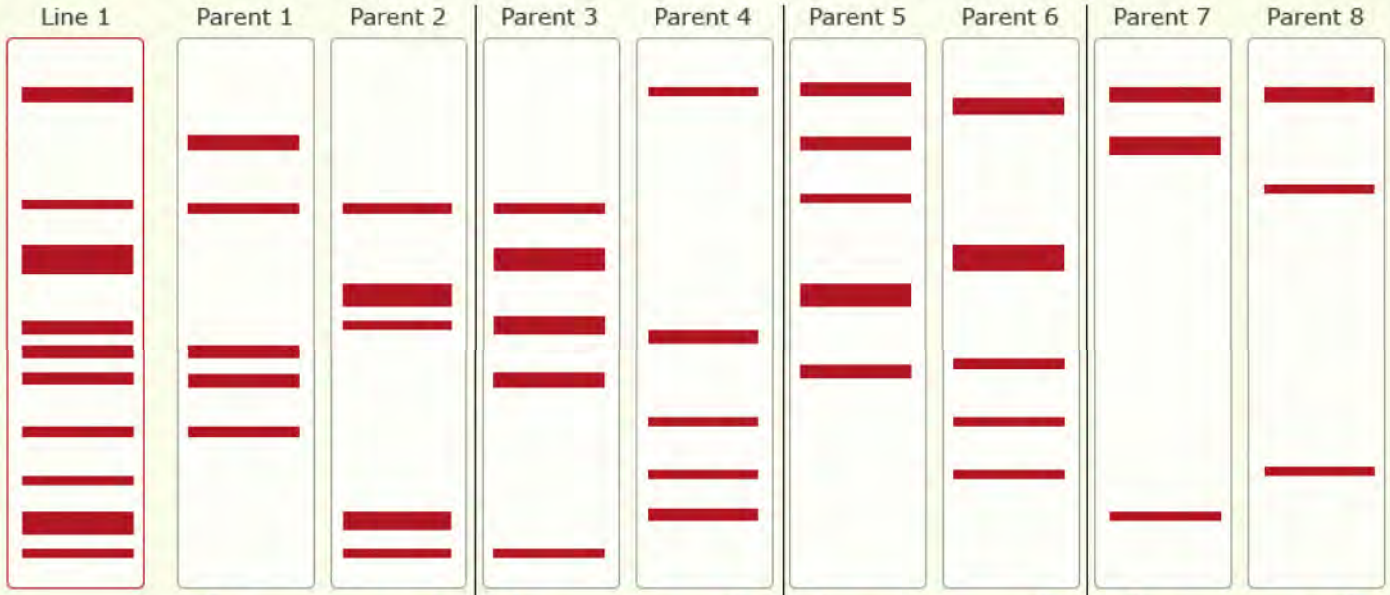


Fig. 21 PCR gel electrophoresis results. Image by Rkalendar. Licensed under Creative Commons Attribution-Share Alike 3.0 via Wikimedia Commons.

TRY THIS!

Genetic Fingerprinting

Look at the DNA fingerprint pattern in Figure 21. Which pair of parent's DNA matches that of Line 1? You can drag the Line 1 banding pattern over the parents' bands to help you decide.



General Application of Markers

GENETIC FINGERPRINTING

The concept of fingerprinting is increasingly being applied to determine the ancestry of plants and animals. Genetic fingerprinting can be used in the breeding of endangered species or commercially important crops because it can help guarantee the authenticity of the plants. With the ability of obtaining highly specific DNA profiles, genetic fingerprinting can be used to protect from illegal use of patented or otherwise registered varieties. For commercially important crops that are difficult to characterize phenotypically, genetic fingerprinting is an important tool to identify genetic diversity within breeding populations. One of the earliest methods of genetic fingerprinting used hybridization-based RLFP markers. An example of genetic fingerprinting data is provided in Figure 22.

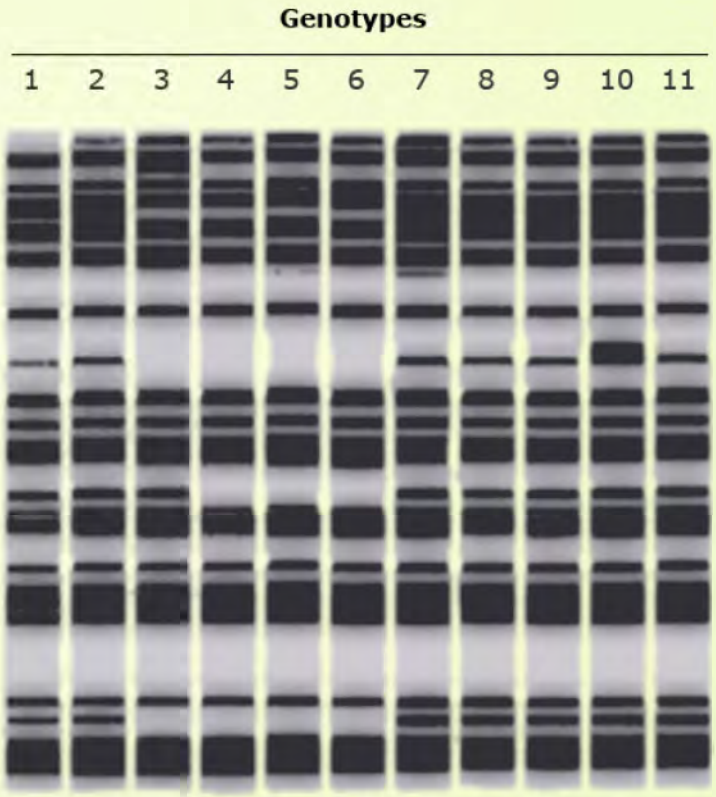


Fig. 22 An example of a genetic fingerprint based on AFLP analysis.

STUDY QUESTIONS

Genetic Fingerprinting

How can genetic similarity for any pair of lines be estimated among the different genotypes in Figure 22?

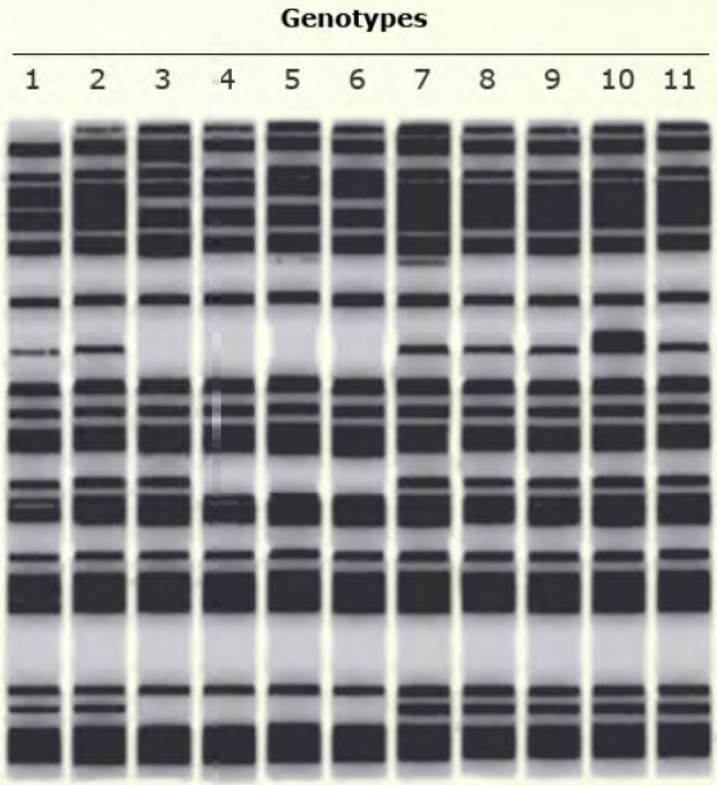


Fig. 22 An example of a genetic fingerprint based on AFLP analysis.

General Application of Markers

GENETIC FINGERPRINTING

Genetic fingerprinting can be applied at all phases of cultivar development. The phases are described at right.

Application of DNA fingerprinting in plants (PDF)

Phase 1: Identifying genetic variation

- In parent selection
- In recurrent selection
- In assigning individuals to heterotic pools
- In choosing genetic resources

Phase 2: Developing variety parents or testing hybrids

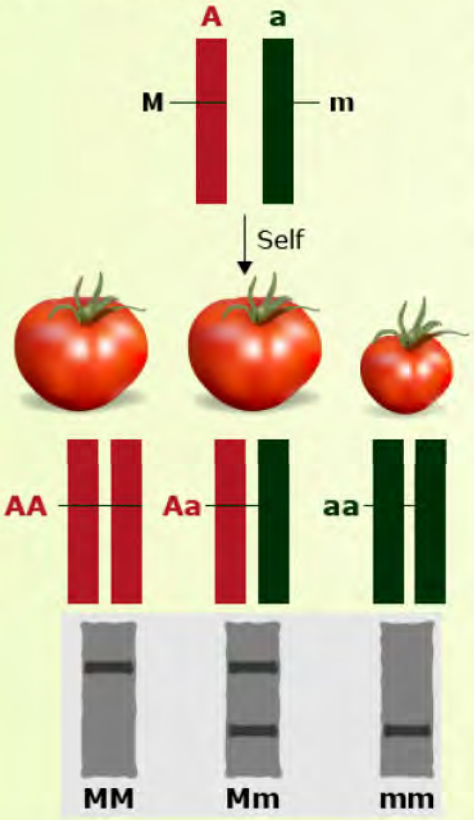
- To measure heterozygosity to predict hybrid performance
- In conducting backcrossing

Phase 3: Seed multiplication and variety protection

- To ensure purity of hybrids and blends
- For variety approval
- To identify "essentially derived varieties" (EDV)

General Application of Markers

GENE TAGGING



If markers flank a gene of interest, the likelihood of a recombination event occurring between markers and gene of interest depends on the genetic distance between them. Thus, the closer the marker is to the gene controlling a trait of interest, the higher chance that there will be no recombination between gene and marker. Absolutely linked markers co-segregate with the trait of interest.

A marker linked to a gene controlling a gene of interest can serve as a "tag" for that gene/trait (Fig. 23).

Fig. 23 An example of gene tagging with a molecular marker completely linked to a trait of interest. A hypothetical gene "A" controls fruit size in tomato and is dominant over "a". A co-dominant marker is available to identify individuals carrying either of the fruit size alleles. The marker "M" is linked to the dominant allele, and "m" to the recessive allele. The detection of markers M and m by PCR produces fragments that can be separated by gel electrophoresis.

STUDY QUESTIONS

General Application of Markers

What might be the benefit(s) of tagging a gene with a molecular marker for a trait that can be phenotypically scored?

General Application of Markers

USE OF LINKED MARKERS AND FINGERPRINTING TO ASSIST BACKCROSSING

Marker-assisted backcrossing involves three steps:

Step 1. Selection of donor allele at the markers linked to target gene to reduce loss of target allele due to recombination. In this step, markers are useful if the trait is controlled by a recessive allele, or when multiple resistance genes are to be obtained from the donor. Also, markers are useful for environmentally-sensitive genes and for expensive phenotypes, for example, grain quality.

Fig. 24 Development of male-sterility by marker-assisted backcrossing in maize. A male sterile donor is crossed with a fertile recurrent parent. Charts depict proportions of donor and recurrent parent genomes; bars depict chromosome segments of donor and recurrent parent. Progeny containing largest proportion of recurrent parent genome can be detected as early as in the BC_1 generation using molecular markers and genetic fingerprinting. Overall, the use of markers helps speed production of new male sterile lines.

General Application of Markers

USE OF LINKED MARKERS AND FINGERPRINTING TO ASSIST BACKCROSSING

Markers are useful for foreground selection of lines having the donor allele in heterozygous condition. An example of the use of markers for foreground selection is described in Figure 25. Without a marker it would not be possible to distinguish progeny heterozygous for the male sterility trait (Msms) from homozygous (MsMs) genotypes because both scenarios result in fertile plants. The use of a co-dominant marker linked to Ms/ms, heterozygotes helps identify heterozygotes and eliminates the need to expend time and resources for selfing and scoring individuals based on pollen production.

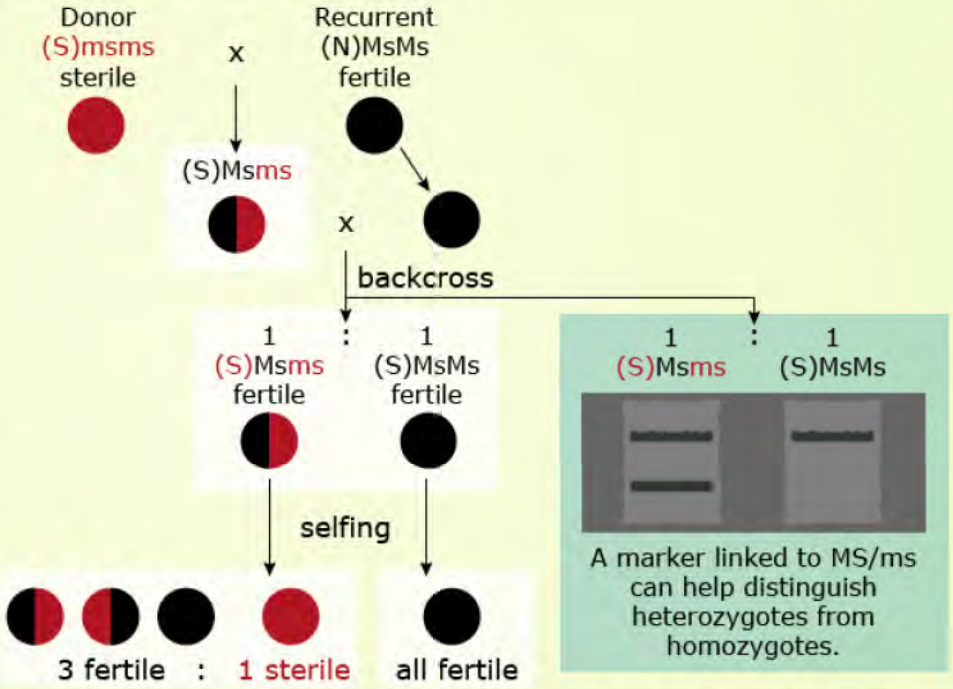


Fig. 25 The use of molecular markers for foreground selection. Backcross of (S)Msms to (N)MsMs produces fertile plants, but of different genotypes (Msms or MsMs). Selfing the MsMs BC_1 progeny will produce all MsMs fertile plants. Selfing of BC_1 Msms progeny will produce fertile and sterile plants in the ratio of 3:1. The use of a linked marker will help eliminate additional work to self and phenotypic screening of the plants.

Non-DNA Markers

Genetic markers are broadly classified into two groups. (1) DNA markers: those based on detection of DNA. (2) Non-DNA markers: those based on visually distinguishable traits, also referred to as morphological markers (e.g. flower color or seed shape); and those based on gene products, referred to as biochemical markers (e.g. RNA, protein, and other cellular metabolites).

The advantage of DNA markers is that they are not affected by environmental factors. However, presence of a particular DNA sequence may not always lead to the expected expression for a trait of interest. This is, because the expression of a particular allele depends on environmental conditions, and also interaction with other genes. Thus, even though an allele with a known effect on a particular trait is present, it might not result in the expected phenotype. Therefore, DNA markers are considered to be a measure of the genetic potential of an individual. The equivalent in human genetics is the risk concept. Based on DNA information, it is possible to predict the risk of a patient for showing a particular condition (e.g., 30% to get pancreatic cancer at a certain age). However, whether this condition is expressed, depends on other circumstances. In contrast, if RNA- or metabolite-based biomarkers for this cancer type are available, onset of this condition can be predicted with high accuracy. Thus, non-DNA markers are indicative of the realized potential of an individual.

Non-DNA Markers

The advantage of morphological markers, also called visible markers, is that they are in general easy to score. However, morphological markers are affected by environmental conditions, making their use less reliable across environments. Also, morphological markers are limited in number compared to the abundance of DNA markers. Biochemical markers are affected by the developmental stage of the plant, and the cell type from which they are isolated. This has to be carefully selected. This is a major difference compared to DNA-markers, which are stable and valid, independent of the tissue from which the respective DNA has been isolated. The World Health Organization defines a biomarker as any parameter that can be used to measure an interaction between a biological system and an environment agent, which may be chemical, physical or biological. Therefore, diagnosis of presence of disease condition and possible treatment requires use of biomarkers. In conclusion, the term biomarker is broadly defined and may include DNA- and non-DNA markers. However, sometimes the term biomarker is used in a more narrow sense for biochemichal non-DNA markers.

Non-DNA Markers

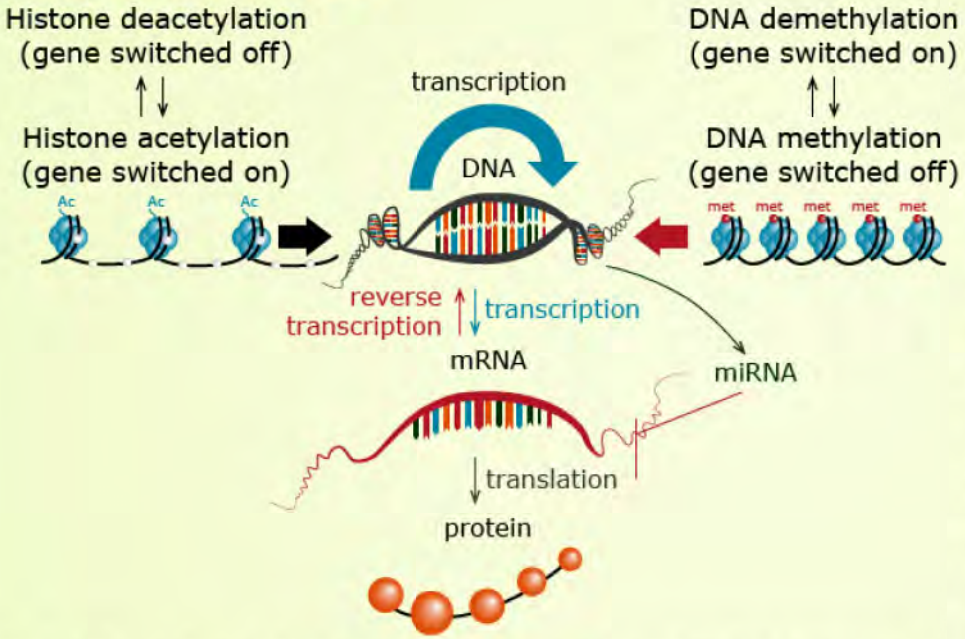


Fig. 26 Scheme of genetic (and epigenetic) information pathways from DNA to RNA to protein. DNA, RNA, and proteins can be used as markers. If the sequence of the protein is known, it may be used to track the DNA (the gene) from which it was encoded. Variation in DNA sequence will result in variation in RNA and protein sequences. If change in the amino acid sequence of an enzyme results in change in its function, the observable phenotype (morphological or biochemical) can be used as a marker. Non-coding RNAs, e.g. miRNA are also important. Many miRNA genes are expressed at specific tissues and developmental stages to regulate expression of specific genes by affecting mRNA stability and translation.

To understand the relationship between DNA markers and non-DNA markers, review the pathways by which genetic information in **deoxyribonucleic acid (DNA)** is transferred to **ribonucleic acid (RNA)** molecules (called transcription), and then transferred from RNA to a protein (termed translation) by a code that specifies the amino acid sequence. Epigenetic mechanisms including DNA methylation/ demethylation and histone acetylation/deacetylation may also impact gene expression (see Fig. 27).

Non-DNA Markers

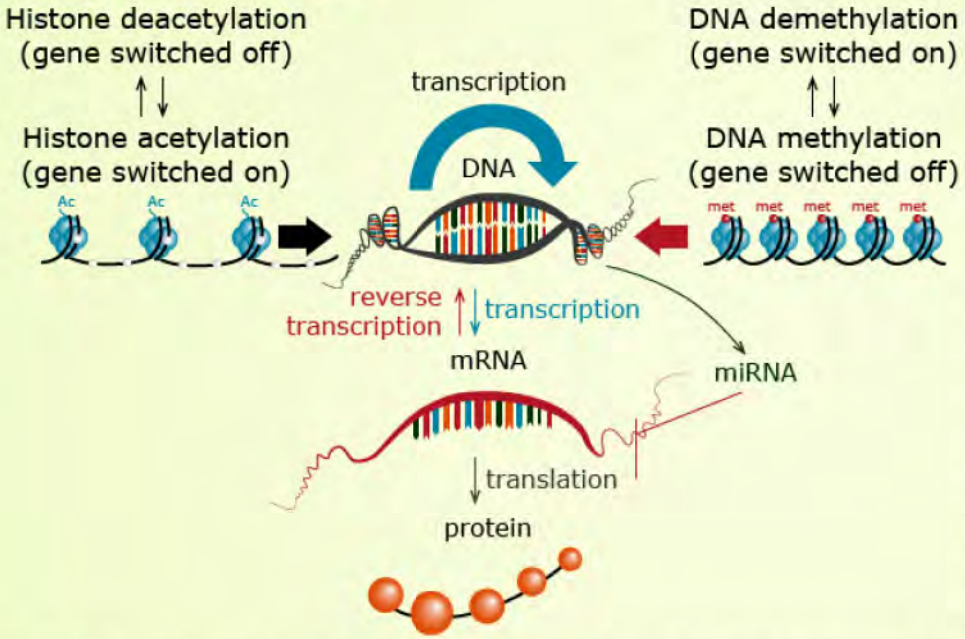


Fig. 26 Scheme of genetic (and epigenetic) information pathways from DNA to RNA to protein. DNA, RNA, and proteins can be used as markers. If the sequence of the protein is known, it may be used to track the DNA (the gene) from which it was encoded. Variation in DNA sequence will result in variation in RNA and protein sequences. If change in the amino acid sequence of an enzyme results in change in its function, the observable phenotype (morphological or biochemical) can be used as a marker. Non-coding RNAs, e.g. miRNA are also important. Many miRNA genes are expressed at specific tissues and developmental stages to regulate expression of specific genes by affecting mRNA stability and translation.

RNA-BASED MARKERS

Figure 26 illustrates two steps in gene expression, transcription (production of mRNA from DNA), and translation (production of proteins from mRNA). Not all genes produce mRNA that can be translated into proteins. Certain genes are transcribed into non-coding RNAs (e.g. micro-RNAs - miRNAs) or short-interfering RNAs - siRNAs) that serve a regulatory function during plant growth and development. A gene can be either "on" or "off" depending on the cell-type, stage of development, and environmental signals, meaning that at any moment each cell makes coding and non-coding RNA from only a proportion of its genome.

Non-DNA Markers

RNA-BASED MARKERS

Microarray Technologies

Development of a microarray starts with the synthesis of probes. Probes can be either (a) cDNA sequences derived from expressed sequence tags (EST) clones or small fragments from PCR or (b) synthetic oligonucleotides, short sequences designed to complement genomic targets of interest. The oligonucleotides may be long (60-mer) or short (25-mer) depending on the purpose of the experiment. Longer probes bind their targets with higher specificity than shorter probes. However, shorter probes may be spotted at a higher density on an array than longer probes, thus reducing the cost of array production.

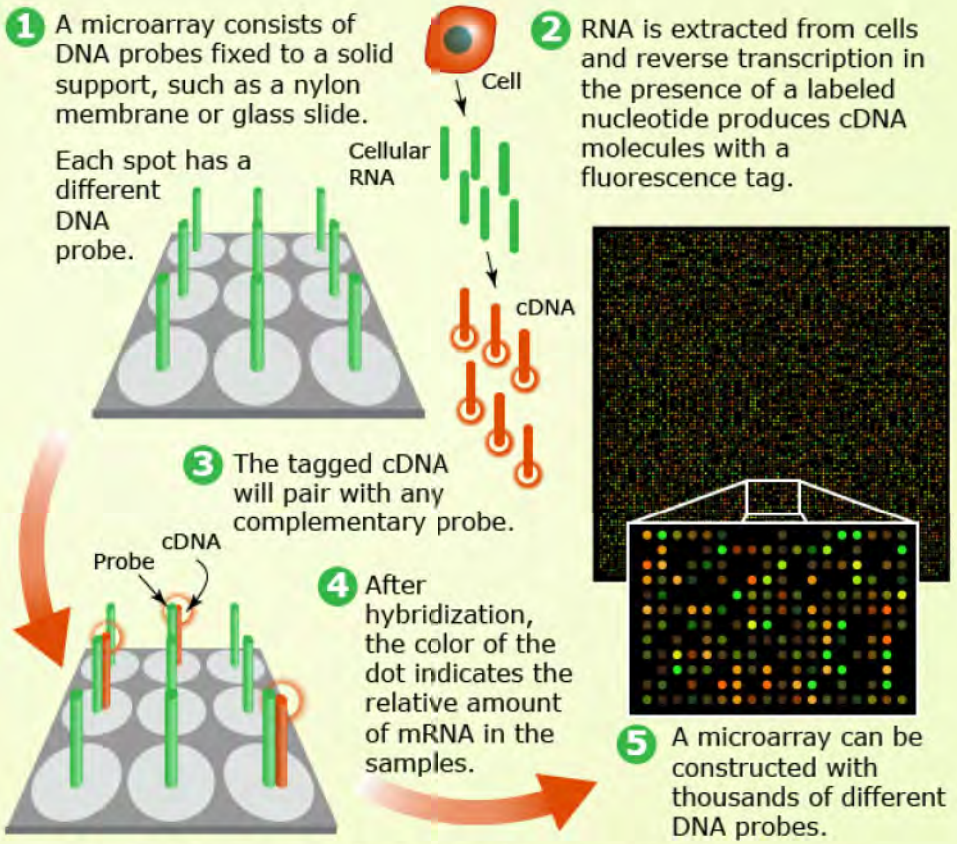


Fig. 27 Microarrays allow the detection of the expression of thousands of genes.

Non-DNA Markers

RNA-BASED MARKERS

Microarray Technologies

Microarray technologies allow parallel assessment of thousands of genes in a single experiment to generate data for gene function, or trait characterization. Microarray analysis involves hybridization of target sequences with gene-specific probes spotted on an array (Fig. 27). For the development of RNA-based markers, target sequences are prepared from total RNA or mRNA.

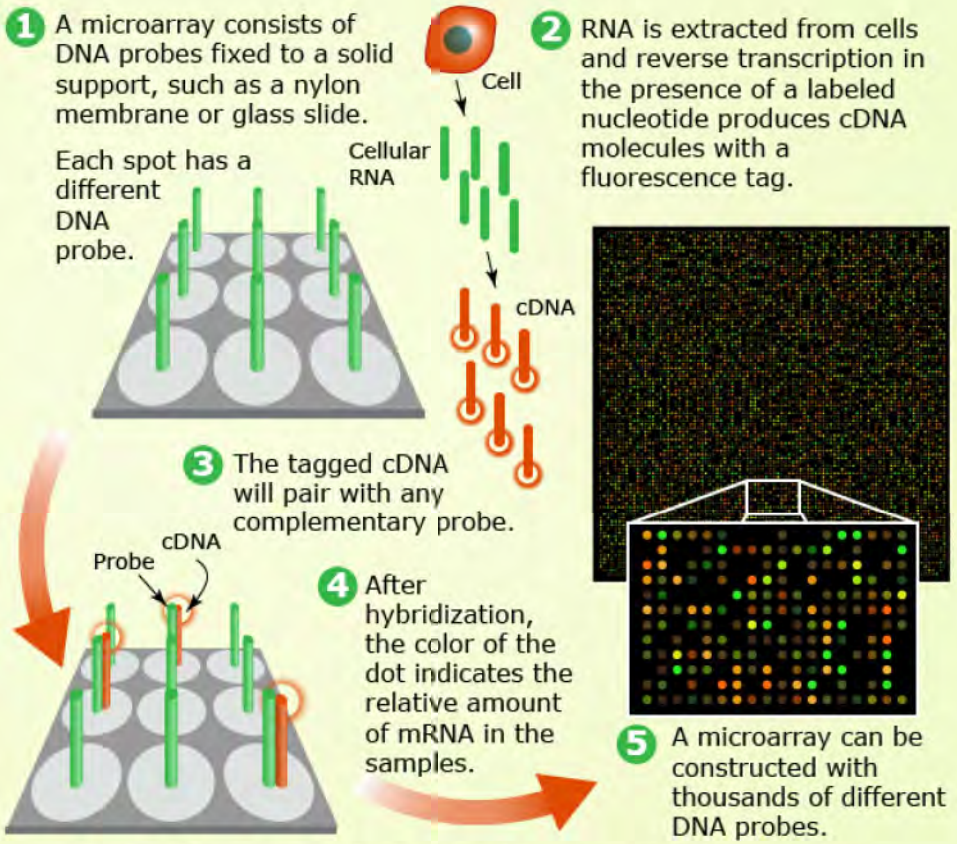


Fig. 27 Microarrays allow the detection of the expression of thousands of genes.

Non-DNA Markers

RNA-BASED MARKERS

Micro RNAs (miRNAs) are small non-coding RNAs which play key roles in regulating the translation and degradation of mRNAs. Genetic or epigentic alterations may affect miRNA expression, thereby leading to aberrant target gene(s) expression in diseases such as cancer. Thus, miRNAs may also provide useful biomarkers for diseases diagnosis. For example, a study by Yanaihara et al. (2006) identified 43 miRNAs that are uniquely expressed in affected lung tissue. Recent studies indicate that miRNAs expression in plants is affected by stress conditions such as drought (Li et al. 2011).

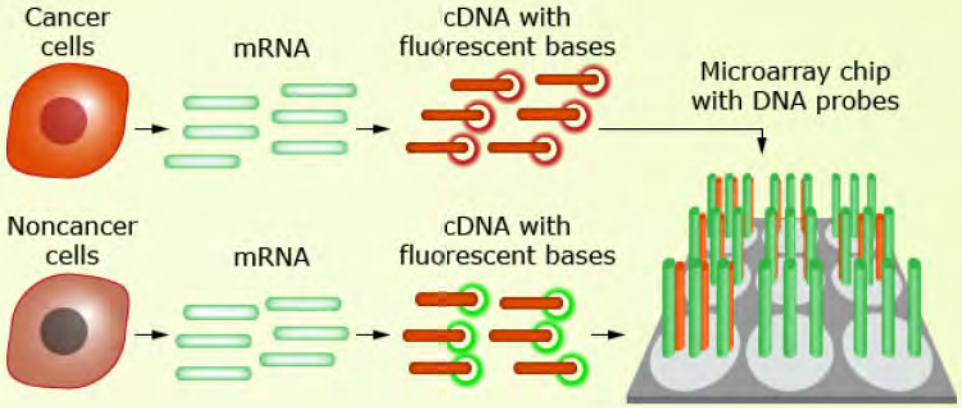


Fig. 28 Detection of variation in gene expression by microarrays to predict the occurrence of cancer.

IN DETAIL

Microarray Analysis

1. On-slide synthesized arrays
Such arrays are prepared by chemical synthesis of probes on the array surface, e.g., Affymetrix arrays (Fig. 29, 30).

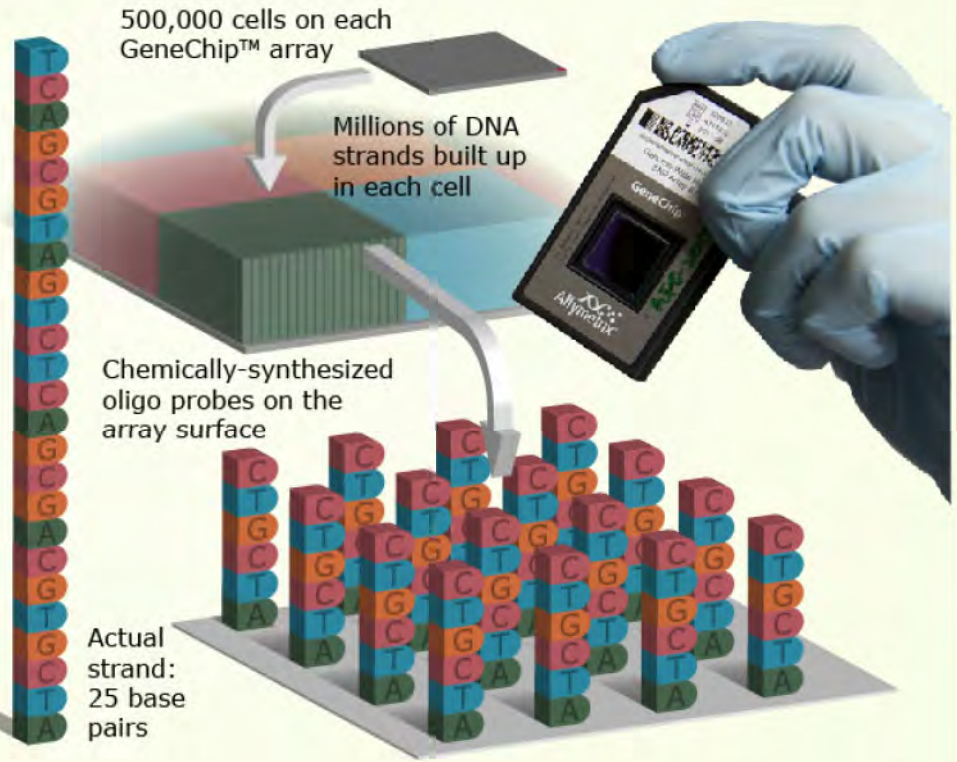


Fig. 29 On-slide synthesized arrays on the Affymetrix GeneChip. The actual size of the chip is 1.28 cm x 1.28 cm and costs about $400. Probe spots on each cell are 10 µm. Oligo probes (25-mers) are synthesized by a chemical process known as photolithographic synthesis. Eleven to twenty "match" probes and 11-20 "mismatch" probes per each gene are spotted on the array surface. There is only one target per each array, and arrays are not reused.

IN DETAIL

Microarray Analysis

1. On-slide synthesized arrays
Such arrays are prepared by chemical synthesis of probes on the array surface, e.g., Affymetrix arrays (Fig. 29, 30).

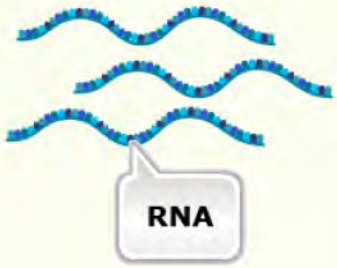


Fig. 30. Analysis of gene expression by the Affymetrix technology. In general, the procedure involves the following steps: (1) isolation of total RNA or mRNA (total RNA will contain coding and non-coding RNAs); (2) preparation of copy-RNA (cRNA) and labeling of cRNA with biotin, and (3) hybridization and detection of cRNA with fluorescently labeled streptavidin.

IN DETAIL

Microarray Analysis

2. Spotted cDNA arrays
This type of arrays is prepared by spotting purified PCR products from a cDNA library on glass using a robotic arrayer.

3. Spotted gene-specific sequence tag arrays
Similar to spotted cDNA arrays, PCR products are spotted on glass by a robotic arrayer. However, in contrast to spotted cDNA arrays, spotted gene-specific sequence tags are developed by PCR using primers targeting unique segments of genes or BAC clones.

4. Spotted long oligonucleotide arrays
These arrays constitute oligos ranging from 50-70 base chemically synthesized to match a particular region of a gene of interest. The 50-70mer oligos are spotted on glass slides robotically.

Fig. 31 The GenePix 4000B Microarray Scanner is used to scan Nimblegen and other arrays spotted in a 1" x 3" format at 5 µm -100 µm resolution (16-bit dynamic range).

Non-DNA Markers

PROTEIN-BASED MARKERS

Common protein markers are **isozymes**. Isozymes are enzymes with similar function derived from more than one locus. Isozymes are encoded by gene families resulting from duplication events. Isozymes are different from allozymes in that allozymes represent one enzyme derived from a single locus.

Isozymes are analyzed by a procedure called electrophoresis. Electrophoresis is a technique for separating macromolecules on a gel by means of an electric field and specific chemical staining (Fig. 32). Therefore, to be useful as markers, isozymes must be electrophoretically resolvable (i.e bands can be clearly separated for visualization on a gel), and detectable by various in-gel assay methods.

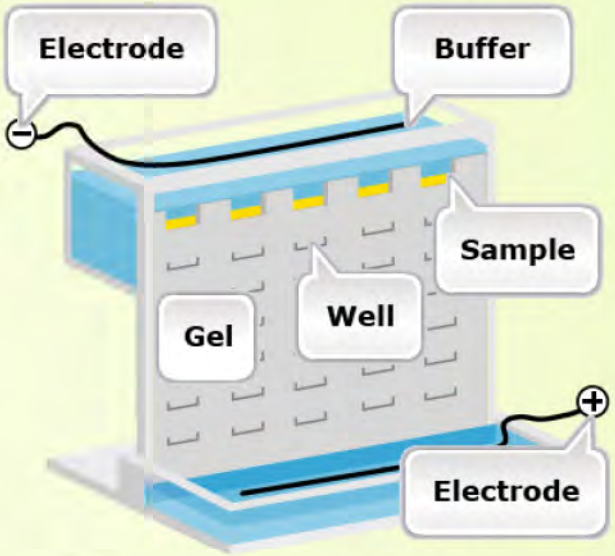


Fig. 32 Electrophoresis is a laboratory technique used to evaluate isozymes.

Non-DNA Markers

PROTEIN-BASED MARKERS

The advantage of isozymes is that they are robust and highly reproducible. Also, isozymes have codominant expression, meaning that both homozygotes can be distinguished from the heterozygote and neither allele is recessive. However isozymes are gene products, so they reveal only a small subset of the actual variation in DNA sequences between individuals and do not reveal variation in the non-coding regions of the genome. Other limitations of isozymes as markers include: (i) data complexity as a result of dimers or multimers of the enzymes; (ii) multi-allelic and multi-locus systems can make interpretation of the banding patterns difficult; (iii) the system is limited to those enzymes that can be detected in situ, resulting in a narrow coverage of the genome; (iv) relatively few biochemical assays are available to detect isozymes; and (vi) the assay is based on a phenotype, and thus sensitive to the environment.

Currently, isozymes are used mainly for germplasm identification and population genetics studies. Other examples of application of proteomic approaches are listed below.

Non-DNA Markers

PROTEIN-BASED MARKERS

Currently, isozymes are used mainly for germplasm identification and population genetics studies. Other examples of application of proteomic approaches are listed below.

1. Two-dimensional polyacrylamide gel electrophoresis was used to detect polymorphic protein markers in several plant species (Vienne et al. 1996) http://www.nature.com/hdy/journal/v76/n2/pdf/hdy199624a.pdf
2. A proteomic approach was used to identify protein markers in lung cancer (Mehan et al. 2012) http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0035157
3. Isozymes were useful in developing the linkage map for tomato (Bernatzky and Tanksley, 1986). http://www.genetics.org/content/112/4/887.full.pdf+html
4. Maquet et al. (1997) used allozyme markers to study the genetic structure of Lima bean (Phaseolus lunatus L.) base collection. http://www.springerlink.com/content/fc5vlymm8nvaqj13/fulltext.pdf
5. Ibáñez et al. (1999) evaluated isozyme uniformity in a wild extinct insular plant [Lysimachia minoricensis J.J. Rodr. (Primulaceae)]. http://onlinelibrary.wiley.com/doi/10.1046/j.1365-294X.1999.00633.x/pdf
6. Rouamba et al. (2001) assessed allozyme variation of onion (Allium cepa L.) populations from West Africa. http://www.springerlink.com/content/rnhg43xtmfp70ald/fulltext.pdf

Non-DNA Markers

METABOLITE-BASED BIOMARKERS

As described earlier, in human health, changes from healthy states to disease state and conditions can be described in terms of important metabolites of cells. A similar approach can be used to determine biomarker metabolites in plants during growth and development. For example, a study by Tarpley et al. (2005) established a biomarker metabolite set for rice during development.

The advantage of metabolite-based markers is that their levels are more closely associated with phenotypes than DNA markers. Therefore, establishing a set of metabolite biomarkers for a plant may be useful in predicting agronomic performance under different environments (Sulpice et al. 2009, 2010; Steinfath et al. 2010).

Non-DNA Markers

METABOLITE-BASED BIOMARKERS

Two techniques used largely for profiling metabolite biomarkers are: (1) mass spectrometry (MS); and (2) nuclear magnetic resonance (NMR). Examples of methods for establishing metabolite biomarkers in plants include gas chromatography-mass spectrometry (GC-MS), liquid chromatography mass-spectrometry (LC-MS), and NMR. The application of some of these methods is described in the work by Skogerson et al. (2009) to establish metabolite profiles for various wines as biomarkers for wine sensory properties. The advantage of such wine biomarkers is that they may be used to replace expensive and laborious sensory panels (Skogerson et al. 2009). Also, such biomarkers may be useful in for various regulatory purposes, for example, detection of adulterations.

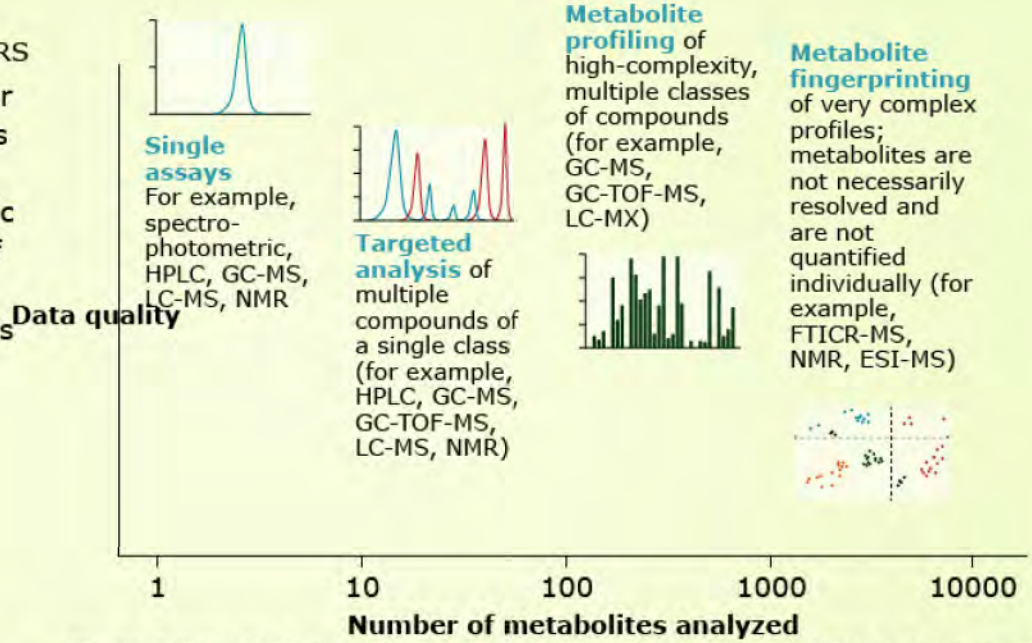


Fig. 33 Relationship between number of metabolites analyzed by MS and NMR techniques and data quality. Analysis of a single compound will result in data of higher quality than analysis of several metabolites in a biochemical pathway, or an entire organism. The current methods are unable to fully cover all metabolites in a cell (metabolomes). Higher plants produce tens of thousands of different metabolites making the analysis challenging. HPLC high performance liquid chromatography, TOF time of flight, FTICR Fourier-transform-ion-cyclotron resonance. Adapted from Fernie et al., 2004.

Non-DNA Markers

METABOLITE-BASED MARKERS

In evaluating biomarkers, there is a trade-off between metabolic coverage and the quality of the metabolite data. As shown in Fig. 33, analysis of a single metabolite or a metabolite class yields data of higher quality than broad analysis for several chemical classes.

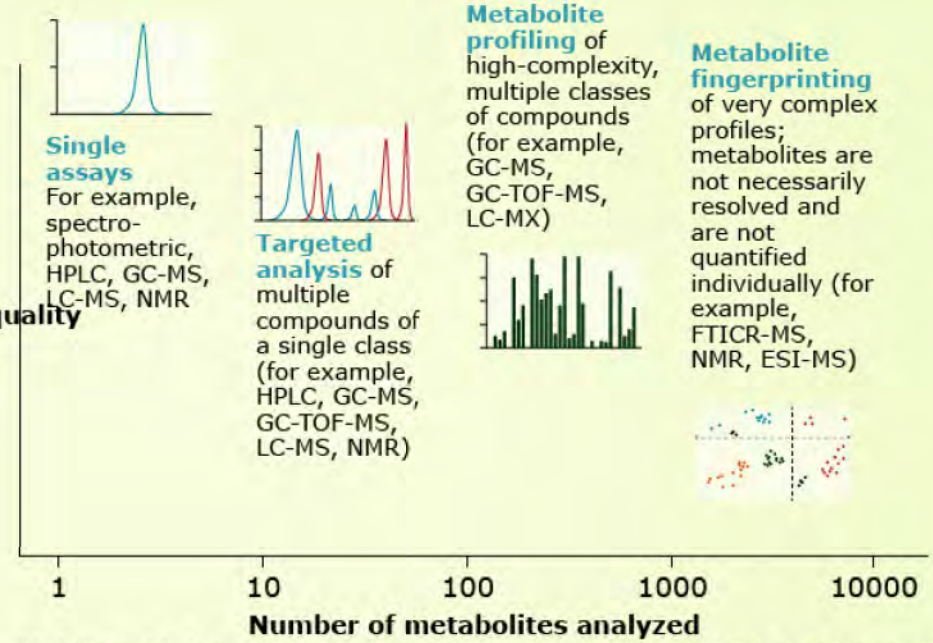


Fig. 33 Relationship between number of metabolites analyzed by MS and NMR techniques and data quality. Analysis of a single compound will result in data of higher quality than analysis of several metabolites in a biochemical pathway, or an entire organism. The current methods are unable to fully cover all metabolites in a cell (metabolomes). Higher plants produce tens of thousands of different metabolites making the analysis challenging. HPLC high performance liquid chromatography, TOF time of flight, FTICR Fourier-transform-ion-cyclotron resonance. Adapted from Fernie et al., 2004.

Non-DNA Markers

PLANT PHENOMICS

Plant phenomics is the study of how genetic makeup of an individual influences its physical and biochemical characteristics in a particular environment (Furbank and Tester, 2011). High-throughput phenomics facilities are using automated plant imaging for the repeated, non-destructive acquisition of high-dimensional phenotypic data on a whole-plant scale.

For example, The Australian Plant Phenomics Facility uses the Plant Accelerator (http://www.plantaccelerator.org.au/).

LemnaTec phenomics systems (http://www.lemnatec.com/) can handle small plants (Arabidopis) and large plants (corn) to measure various parameters including, leaf area, chlorophyll content, stem diameter, height, biomass, color, and leaf tracking over time. The Lemna Tec technology can also be used to measure responses to salt, and drought stress. The application of phenomics is important for studying complex stress traits such as drought because non-destructive imaging methods allow temporal resolution and monitoring of the same plants during the experiment (Berger et al. 2010).

Ultimately, phenomic data (e.g., canopy reflectance) can be used as indirect trait for agronomic traits of interest. An example is the measurement of canopy "greenness" to describe the nitrogen use efficiency (NUE) of plant genotypes (https://www.pioneer.com/home/site/us/agronomy/library/template.CONTENT/guid.8AA5E524-D466-6643-809B-DA3586758BEA#canopy).

FURTHER THOUGHT

Discussion

Non-DNA biomarkers are important diagnostic tools in human genetics, but not in plant breeding so far. Discuss possible reasons, and also applications that you see for non-DNA biomarkers in plants in future.

FURTHER THOUGHT

Reflection

The Module Reflection appears as the last "task" in each module. The purpose of the Reflection is to enhance your learning and information retention. The questions are designed to help you reflect on the module and obtain instructor feedback on your learning. Submit your answers to the following questions to your instructor.

1. In your own words, write a short summary (< 150 words) for this module.

2. What is the most valuable concept that you learned from the module? Why is this concept valuable to you?

3. What concepts in the module are still unclear/the least clear to you?

References

Akhunov, E., C. Nicolet, J. Dvorak. 2009. Single nucleotide polymorphism genotyping in polyploidy wheat with the Illumina GoldenGate assay. Theor Appl Genet 119:507-517

Baskin et al. 2009. http://www.geospiza.com/Products/WhitePaper_06102009.pdf

Clark, R.M. 2010. Genome-wide association studies coming of age in rice. Nature Genetics 42:11, 926-927.

Craig, D.W., et al. 2008. Identification of genetic variants using bar-coded multiplexed sequencing. Nature Methods 5, 887-893. doi:10.1038/nmeth.1251

DiGuistini et al. 2009. De novo genome sequence assembly of a filamentous fungus using Sanger, 454 and Illumina sequence data. Genome Biology 10:R94
http://www.biomedcentral.com/content/pdf/gb-2009-10-9-r94.pdf

Eid et al. 2009. Real-Time DNA sequencing from single polymerase molecules. Science 323: 133-138http://www.sciencemag.org/content/323/5910/133.full.pdf

Elshire et al. 2011. PLoS ONE 6: doi:10.1371/journal.pone.0019379
http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0019379

Fernie et al. 2004. Metabolite profiling: from diagnostics to systems biology. Nature Reviews Molecular Cell Biology 5, 763-769 (September 2004) | doi:10.1038/nrm1451

References

Flicek, P., and E. Birney. 2009. Sense from sequence reads: methods for alignment and assembly. Nature methods. 6:S6-S12. http://www.nature.com/nmeth/journal/v6/n11s/pdf/nmeth.1376.pdf

Hawkins et al. (2010) Next-generation genomics: an integrative approach. Nature Reviews Genetics. 11:476-486. http://www.nature.com/nrg/journal/v11/n7/pdf/nrg2795.pdf

Huang et al. 2009. High-throughput genotyping by whole-genome resequencing. Genome Res. 2009 Jun;19(6):1068-76. doi: 10.1101/gr.089516.108.

Illumina, Inc. 2006. GoldenGate Assay workflow. https://www.illumina.com/documents/products/workflows/workflow_goldengate_assay.pdf

Li, H., and N. Homer. 2010. A survey of sequence alignment algorithms for next-generation sequenc-ing. Briefings in Bioinformatics. 11: 473-483. http://bib.oxfordjournals.org/content/11/5/473.full.pdf+html

MaizeGDB. Maize bin viewer. http://www.maizegdb.org/bin_viewer

Maxam, A M., and W. Gilbert. 1977. A new method for sequencing DNA. Proc. Natl. Acad, Sci. USA. 74:560-564. http://www.pnas.org/content/74/2/560.full.pdf

Metzker (2010) Sequencing technologies — the next generation. Nature Reviews Genetics. 11:31-46] is ideal for covering the Objective 3. http://www.nature.com/nrg/journal/v11/n1/pdf/nrg2626.pdf

References

Munroe, D., and T. J. R. Harris. 2010. Third-generation sequencing fireworks at Marco Island. Nature Biotechnol 28:426-428 http://www.nature.com/nbt/journal/v28/n5/pdf/nbt0510-426.pdf

Perkel, J. 2008. SNP genotyping: six technologies that keyed a revolution. Nature Methods 5:447-454

Poland, J.A., P.J. Brown, M.E. Sorrells, J-L Jannink. 2012. Development of High-Density Genetic Maps for Barley and Wheat Using a Novel Two-Enzyme Genotyping-by-Sequencing Approach. PLoS ONE 7(2): e32253. doi:10.1371/journal.pone.0032253

Rothberg et al. 2011. An integrated semiconductor device enabling non-optical genome sequencing. Nature. 475:348-352. http://www.nature.com/nature/journal/v475/n7356/pdf/nature10242.pdf

Salathia, N., H. N. Lee, T. A. Sangster et al. 2007. Indel arrays: an affordable alternative for genotyping. Plant J. 51: 727-737

Schneeberger et al. 2009. SHOREmap: simultaneous mapping and mutation identification by deep sequencing. Nat Methods. 6:550-551.
http://www.nature.com/nmeth/journal/v6/n8/pdf/nmeth0809-550.pdf

References

Schneeberger, K., and D. Weigel. 2011. Fast-forward genetics enabled by new sequencing technologies. Trends Plant Sci. 16:282-8. http://ac.els-cdn.com/S136013851100029X/1-s2.0-S136013851100029X-main.pdf?_tid=69c29f8b83bfc51c21438db4d67728a3&acdnat=1334590816_547510fc0545ec6df921fc0017404666

Shendure, J., and H, Ji.2008. Next-generation DNA sequencing. Nature Biotechnology 26:1135-1145. http://www.nature.com/nbt/journal/v26/n10/pdf/nbt1486.pdf

Syvänen, A. 2001. Accessing genetic variation: genotyping single nucleotide polymorphisms. Nat Rev Genet. 2:930-942

Syvänen, A. 2005. Toward genome-wide SNP genotyping. Nat Genet. 37: S5-S10

ThermoFisher Scientific. Contract Research Organizations To Adopt Ion Torrent Next-Generation Sequencing Platform. http://news.thermofisher.com/press-release/life-technologies/contract-research-organizations-adopt-ion-torrent-next-generation--0

Xu, Y. Molecular Plant Breeding. CABI, Wallingford, Oxon.

Molecular Plant Breeding

Markers and Sequencing

This module was developed as part of the Bill & Melinda Gates Foundation Contract No. 24576 for Plant Breeding E-Learning in Africa.

Funding provided by:

BILL & MELINDA GATES foundation

Other collaborating organizations:

AGRA Growing Africa's Agriculture

CGIAR

Partnering universities:

IOWA STATE UNIVERSITY OF SCIENCE AND TECHNOLOGY

Molecular Plant Breeding **Module 2 Co-authors:**
Thomas Lübberstedt, Madan Bhattacharyya, Walter Suza *(ISU)*

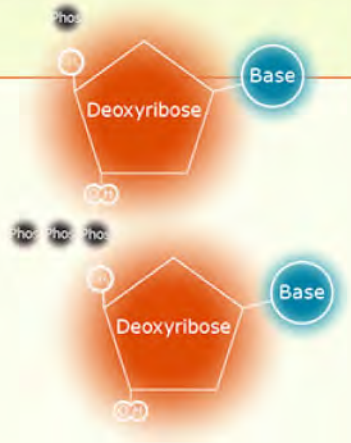
Multimedia Developers:
Gretchen Anderson, Todd Hartnell, and Andy Rohrback *(ISU)*

Molecular Plant Breeding **Course Team:**
William Beavis, Thomas Lübberstedt, Ursula Frei, Walter Suza *(ISU)*; Richard Akromah *(KNUST)*; Richard Edema *(MAK)*; John Derera *(UKZN)*; Ndeye Ndack Diop, Mark Sawkins *(IBP GCP)*

FYI

Dideoxy Sequencing

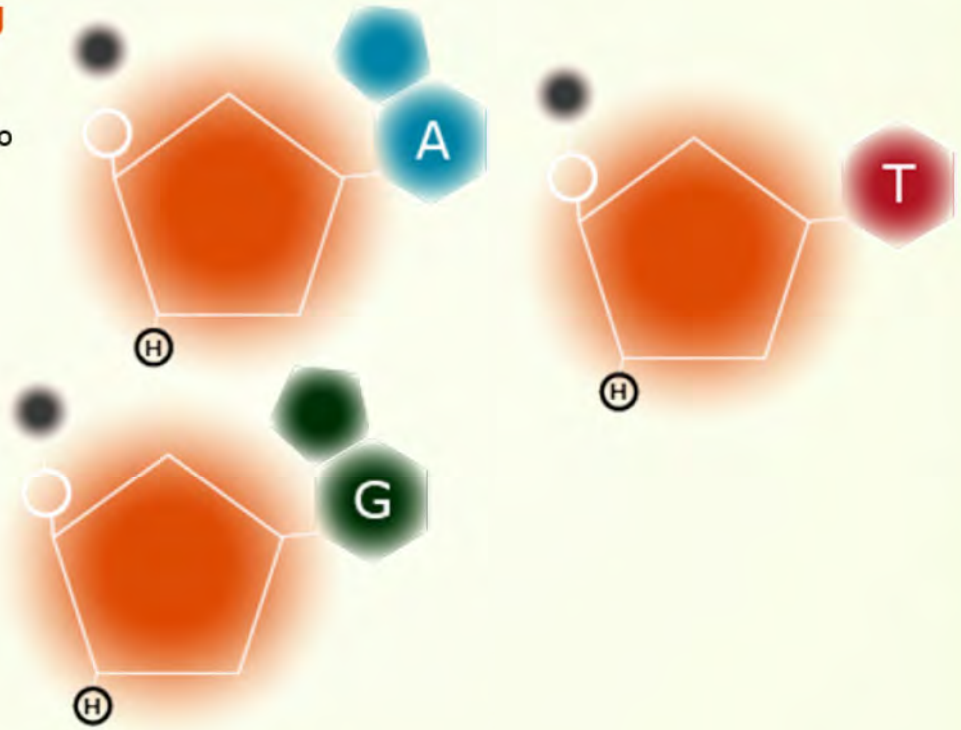
Nucleotides form chains when the hydroxyl part of the phosphate group of one nucleotide bonds with the hydroxyl part of the sugar moiety in the preceding nucleotide.



FYI

Dideoxy Sequencing

The Sanger method uses dideoxynucleotides of each type of base (A, T, C and G) to break the DNA chain.



FYI

Dideoxy Sequencing

[illegible] DNA [illegible] works. [illegible] template [illegible] have been [illegible] with dideoxynucleotide ddCTP, which will halt the sequencing when added to the chain. This can result in three possible sequencing outcomes.

FYI

Dideoxy Sequencing

The ddCTP nucleotide may not be added at all. In that case, the entire sequence would be returned.

FYI

Dideoxy Sequencing

The ddCTP nucleotide may be added at the second location for a C. That would cut the sequence short by one nucleotide.

FYI

Dideoxy Sequencing

The ddCTP nucleotide may be added at the first C location. That would return an even shorter sequence.

FYI

Dideoxy Sequencing

Thus, there are three possible DNA chains that could be returned from the ddCTP process.

FYI

Dideoxy Sequencing

The procedure is repeated with dideoxynucleotide ddGTP, which will halt the sequencing when added to the chain at a G position. This can result in two possible sequencing outcomes.

FYI

Dideoxy Sequencing

The ddGTP nucleotide may not be added at all. In that case, again, the entire sequence would be returned.

FYI

Dideoxy Sequencing

The ddGTP nucleotide may be added at the location for a G nucleotide. That would cut the sequence short by two nucleotides.

FYI

Dideoxy Sequencing

Thus, chains of four nucleotides produced by the ddGTP process indicate the presence of a G nucleotide.

FYI

Dideoxy Sequencing

DNA chains sequenced in the presence of dideoxynucleotide ddATP will have their sequencing halted at A positions. This can result in two possible sequencing outcomes.

FYI

Dideoxy Sequencing

The ddATP nucleotide may not be added at all. In that case, as in all other processes, the entire sequence would be returned.

FYI

Dideoxy Sequencing

The ddATP nucleotide may be added at the location for an A nucleotide. That would cut the sequence down to only one nucleotide.

FYI

Dideoxy Sequencing

Thus, chains of a single nucleotide produced by the ddATP process indicate the presence of a A nucleotide.

FYI

Dideoxy Sequencing

DNA chains sequenced in the presence of dideoxynucleotide ddTTP will have their sequencing halted at T positions. This can result in two possible sequencing outcomes.

FYI

Dideoxy Sequencing

The ddTTP nucleotide may not be added at all. In that case, as in all other processes, the entire sequence would be returned.

FYI

Dideoxy Sequencing

The ddTTP nucleotide may be added in the second location for a T nucleotide. However, that sequence would be the same length as the full sequence.

FYI

Dideoxy Sequencing

The ddTTP nucleotide may be added in the first location for a T, returning a sequence of two nucleotides.

FYI

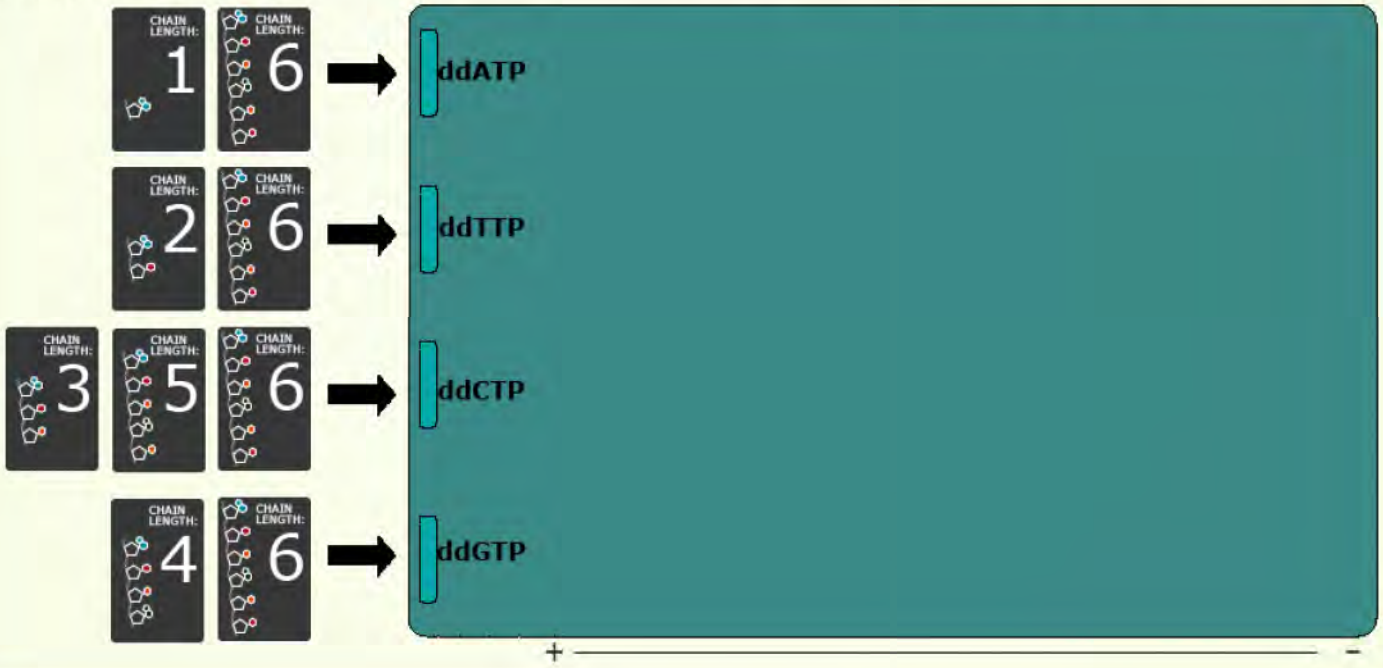
Dideoxy Sequencing

Thus, there are three possible DNA chains that could be returned from the ddTTP process.

FYI

Dideoxy Sequencing

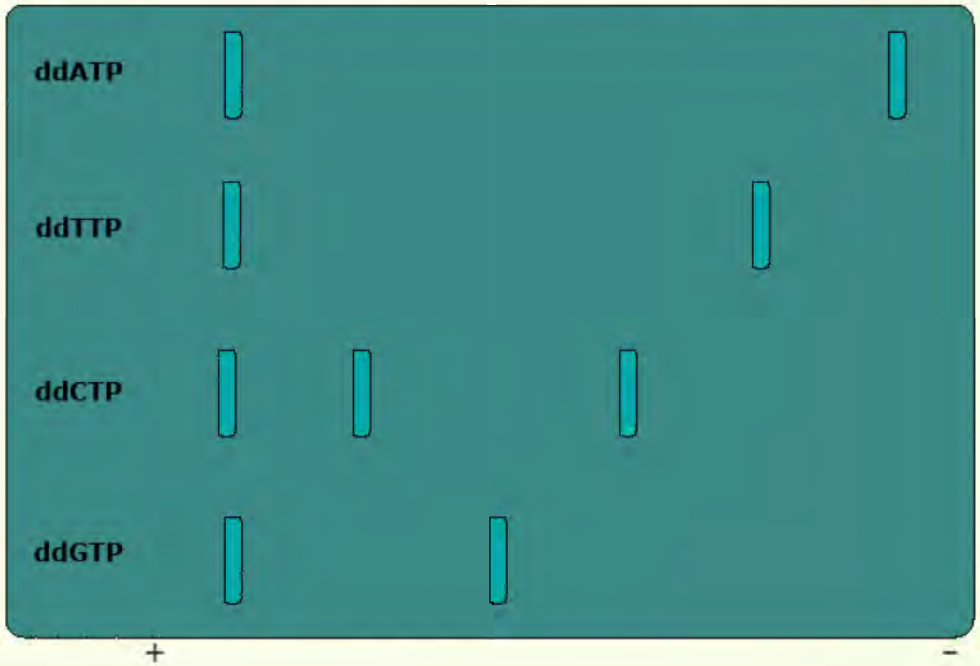
The chains produced by the ddNTP reactions are put through gel electrophoresis to determine their length. Longer chains will move more slowly; shorter chains will move further in the same amount of time.



FYI

Dideoxy Sequencing

The chains produced by the ddNTP reactions are put through gel electrophoresis to determine their lengths. Longer chains will move more slowly; shorter chains will move further in the same amount of time.



FYI

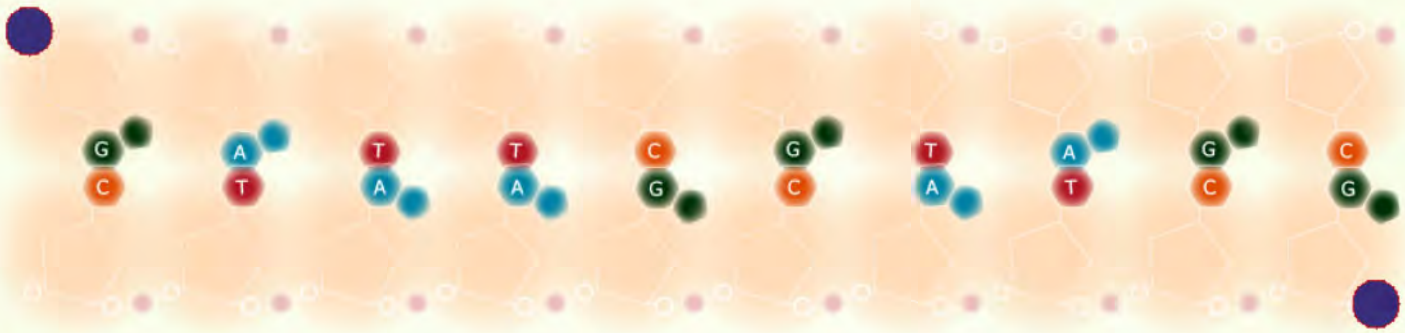
Dideoxy Sequencing

Now that we have the nucleotides in the assembled strand, we can use their complements to understand which nucleotides are in the original template strand.

FYI

Maxam and Gilbert DNA Sequencing

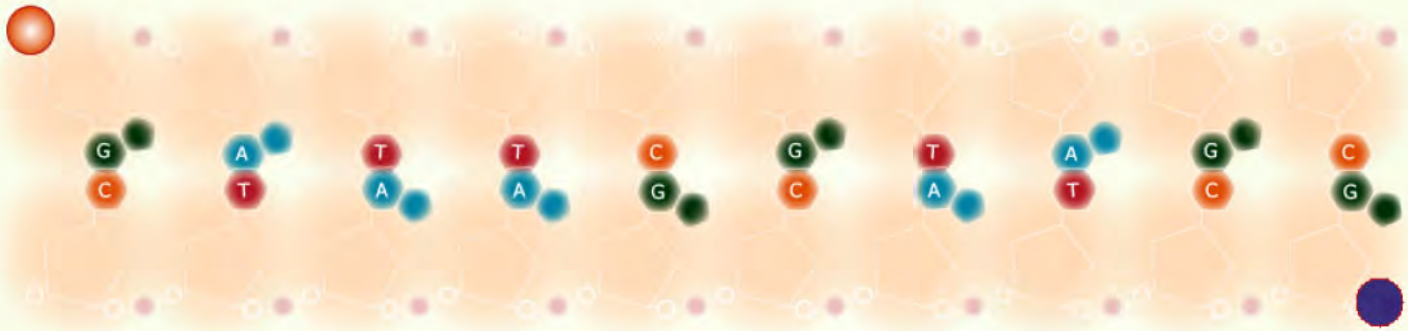
First, the strands are labeled at the 5′ end with a ^{32}P radioactive marker.



FYI

Maxam and Gilbert DNA Sequencing

First, the strands are labeled at the 5′ end with a ^{32}P radioactive marker.



FYI

Maxam and Gilbert DNA Sequencing

A solution containing the labeled single-stranded DNA is distributed into four separate tubes.

FYI

Maxam and Gilbert DNA Sequencing

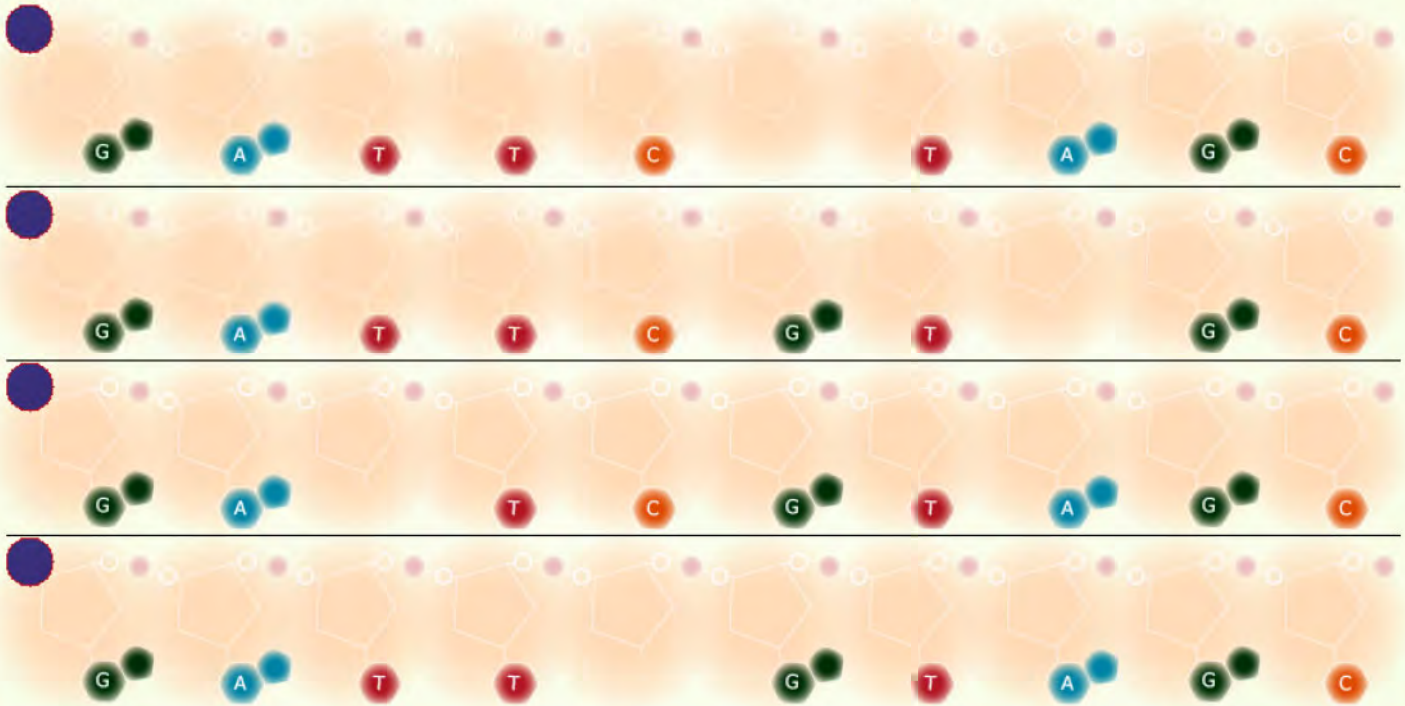
Specific chemical agents added to modify specific DNA bases and cleave the strand into fragments of varying lengths.

FYI

Maxam and Gilbert DNA Sequencing

Each chemical reagent destroys one or two DNA bases.

Experimental conditions are set so that only a few of the bases have been destroyed. It is important to control the conditions of the reaction such that only a few sites are cleaved per DNA molecule.

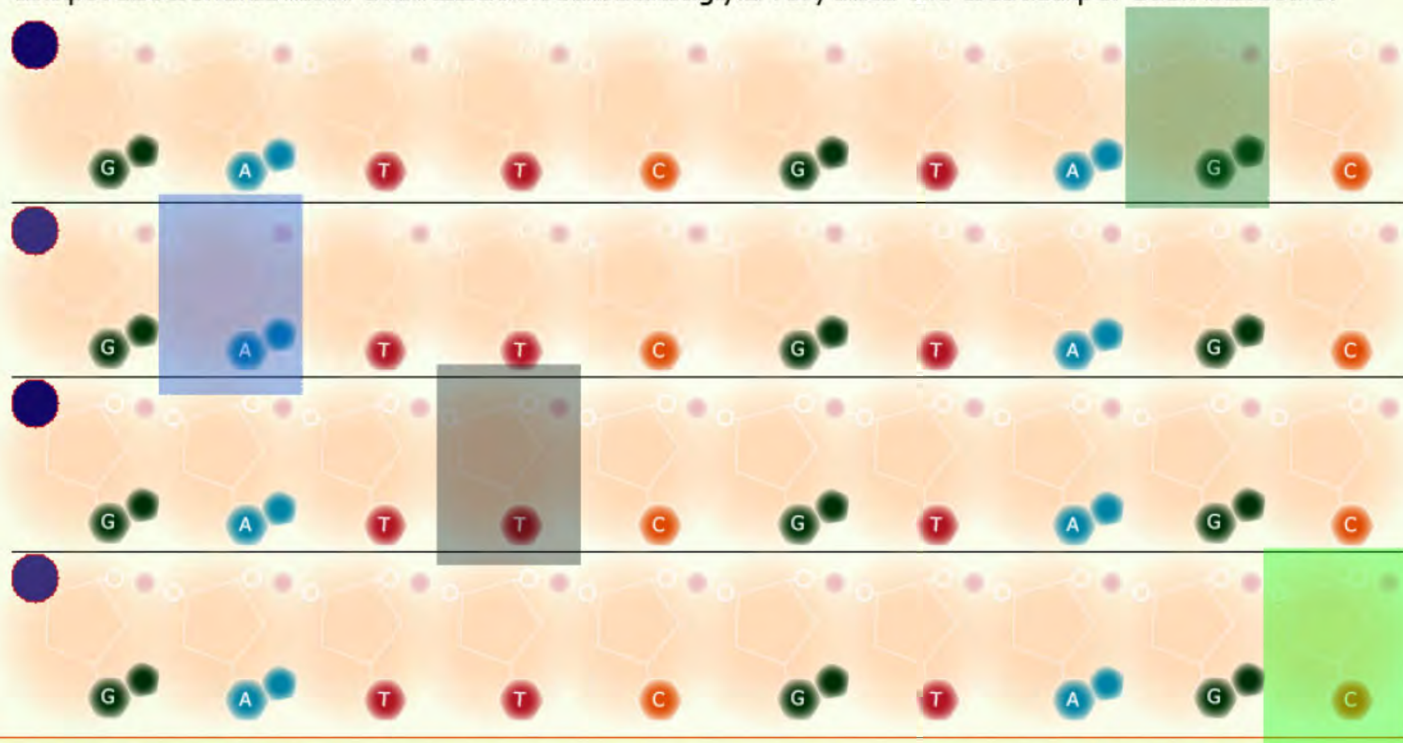


FYI

Maxam and Gilbert DNA Sequencing

The chemicals will cleave each strand of DNA at a different location for the targeted bases. This will produce fragments whose lengths vary with the location of each base.

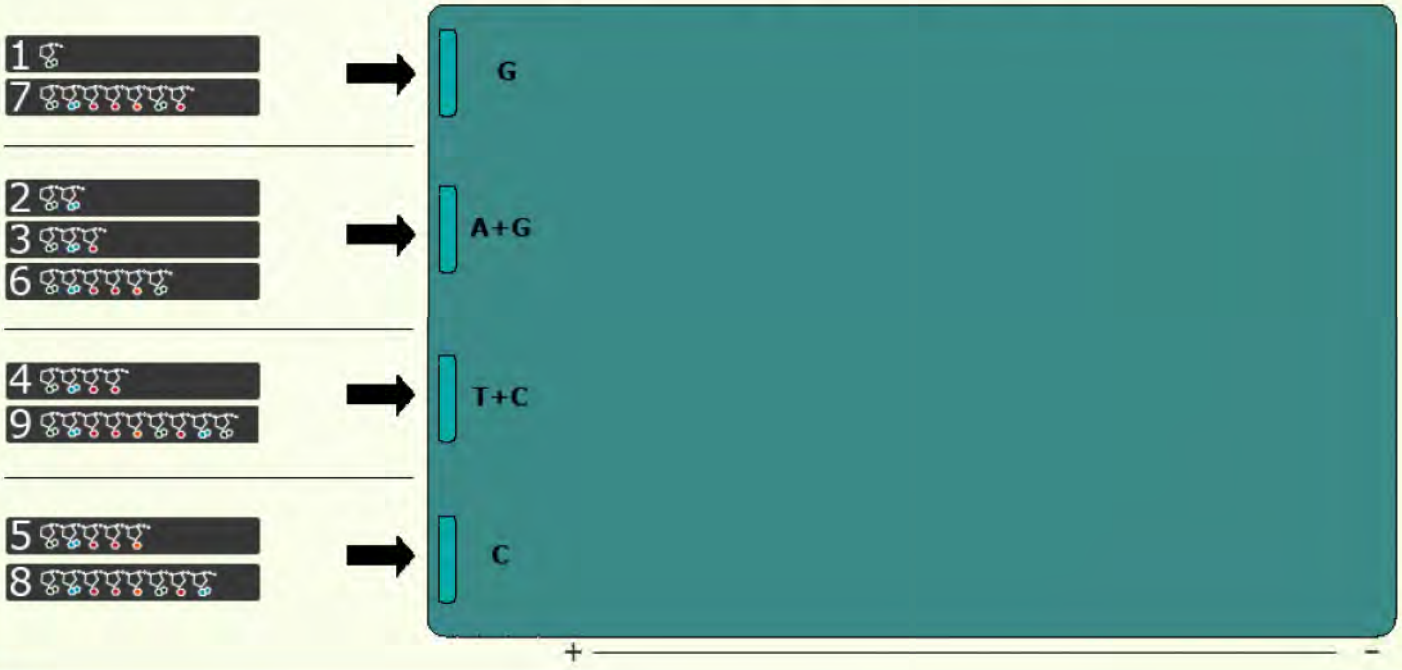
Pretreatments will damage one of the DNA bases which is then destroyed. It is important to control the conditions of each reaction such that only a few sites are cleaved per DNA molecule.



FYI

Maxam and Gilbert DNA Sequencing

The products of the chemical reactions are evaluated by gel electrophoresis and autoradiography. Bands correspond to fragments obtained from cleaving each damaged base.



FYI

Maxam and Gilbert DNA Sequencing

The procedure of the chemical sequencing method is as follows: First, the DNA is labeled at the 5' end with a radioactive phosphate group. Then, the DNA is cleaved into four separate reactions, each of which cleaves the DNA at a specific base. The cleavage reactions are carried out under conditions such that, on average, each DNA molecule is cleaved only once. The products of the four reactions are then loaded into four adjacent wells of a polyacrylamide gel and separated by electrophoresis. Each reaction produces a set of single-stranded DNA fragments that terminate at the base that was cleaved. Bands corresponding to fragments terminating at a particular base are detected by autoradiography, revealing which base is present.