# TOPIC O: ISOLATING, CLONING AND SEQUENCING DNA

## Learning Outcome

*Applications Topic 6*

Candidates should be able to:

(a) Describe the natural function of restriction enzymes.

(b) Explain the formation of recombinant DNA molecule.

(c) Outline the procedures for cloning a eukaryotic gene in a bacterial plasmid and describe the properties of plasmids that allow them to be used as DNA cloning vectors.

(d) Distinguish between a genomic DNA and cDNA library. (Outline of the process of the formation of the libraries and applications of each of the types of library is required.)

(e) Explain how eukaryotic genes are cloned using *E. coli* cells to produce eukaryotic proteins to avoid the problems associated with introns.

(f) Outline two important proteins that can be produced by genetic engineering technique (e.g. human growth hormone and insulin).

(g) Describe the polymerase chain reaction (PCR) and explain the advantages and limitations of this procedure.

(h) Explain how gel electrophoresis is used to analyse DNA.

(i) Outline the process of nucleic acid hybridization and explain how it can be used to detect and analyse restriction fragment length polymorphism (RFLP).

(j) Explain how RFLP analysis facilitated the process of :

  i. genomic mapping in terms of linkage mapping;

  ii. diseases detection, e.g. sickle cell anaemia;

  iii. DNA fingerprinting.

  (Details of application of gel electrophoresis, PCR and nucleic acid hybridisation in RFLP may be required.)

(k) Discuss the goals and implications of the Human Genome Project, including the benefits and difficult ethical concerns for humans. (Knowledge of the technical procedure of the Human Genome Project and DNA sequencing is **not** required.)

## Content Outline

1. Introduction
2. Basic Molecular Techniques in Genetic Engineering
   (a) Restriction and Ligation
   (b) Addition of sticky ends to DNA molecules with blunt ends
   (c) Polymerase Chain Reaction
   (d) Gel Electrophoresis
   (e) Nucleic Acid Hybridisation
3. DNA Libraries
   (a) Genomic Library
   (b) Complementary DNA Library
4. Gene Cloning
   (a) Plasmids
   (b) Process of Gene Cloning
   (c) Production of Proteins using Genetic Engineering
   (d) Other Products of Genetic Engineering
5. Restriction Fragment Length Polymorphism Analysis
   (a) Disease Detection
   (b) DNA Fingerprinting
   (c) Genome Mapping
6. Human Genome Project
   (a) Goals
   (b) Benefits
   (c) Ethical Issues

## References

1. Brooker, R. J. (2005) Genetics: Analysis and Principles. Chapter 18: Recombinant DNA Technology. Second Edition. McGraw-Hill.
2. Campbell, N. A. and Reece J. B. (2011) Biology. Chapter 20: Biotechnology. Ninth Edition. Pearson Education. Inc.
3. Clegg, C. J. and MacKean, D. G. (2000) Advanced Biology: Principles and Applications. Chapter 29: Applications of genetics. Second Edition. John Murray (Publishers) Ltd.
4. Nicholl, D. S. T. (1994) An Introduction to Genetic Engineering. Cambridge University Press.
5. Weaver R. F. (2005) Molecular Biology, Chapter 24: Genomics and Proteomics. Third Edition. McGraw-Hill International Edition.

## 1. Introduction

**Biotechnology** (technology based on biology) is the manipulation of living systems and organisms to make and develop useful products. It includes early practices such as selective breeding of farm animals and using microorganisms to make wine and cheese.

Encompassed within biotechnology is **genetic engineering**, which is the direct manipulation of genes for practical purposes. Applications of genetic engineering include the manufacture of protein products, such as hormones.

The sequencing of the human genome was largely completed by 2003.The potential benefits of the Human Genome Project are numerous yet there are ethical and moral implications behind many of the benefits.

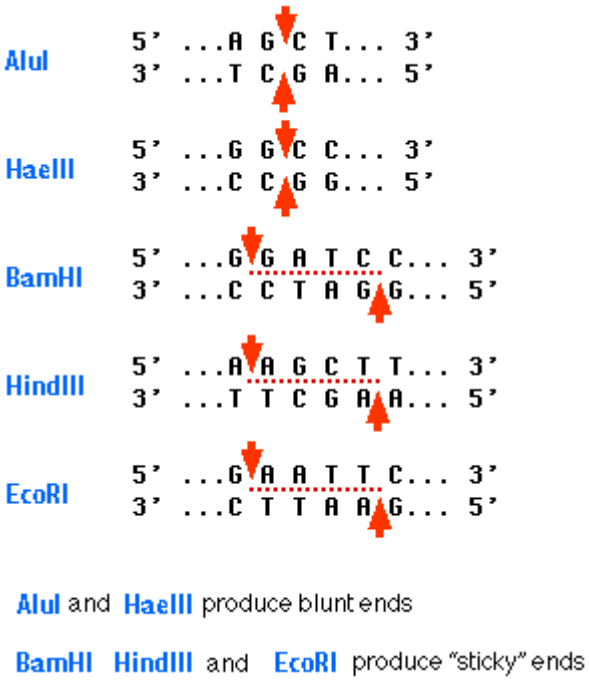
This topic covers the basic molecular techniques required for genetic engineering and how cloning of useful genes can lead to production of human proteins using bacteria cells. The goals, benefits and ethical issues surrounding the Human Genome Project will also be discussed.

## 2. Basic Molecular Techniques in Genetic Engineering

### (a) Restriction and Ligation

(i) **Restriction Enzymes** (also called restriction endonucleases)

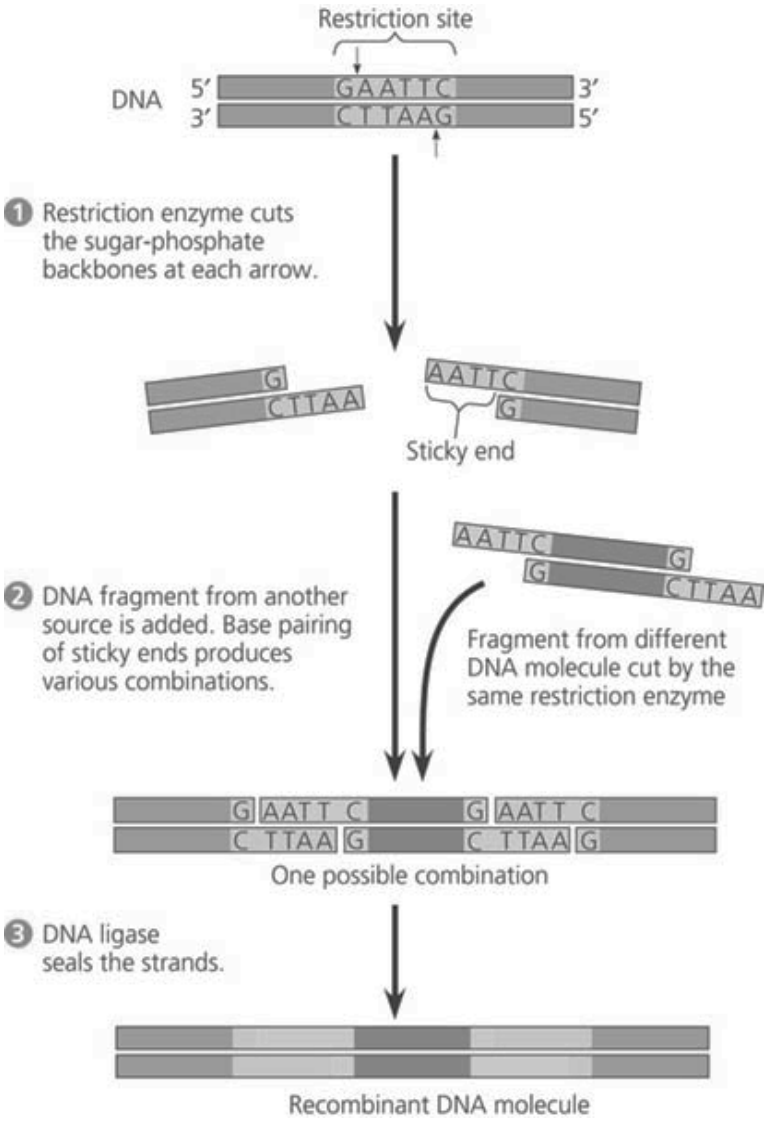
- Naturally found in bacteria.
- **Protect the bacterial cells** from other organisms or phages by **cutting up foreign DNA**, a process known as restriction.
- DNA of a bacterial cell is **protected from the cell's own restriction enzymes**
  - **Addition of methyl groups (-$CH_3$) to adenines or cytosines** within the sequences recognised by the enzymes.
- **Very specific** in action
  - Recognises a short specific DNA sequence known as the **restriction site** (usually four to eight nucleotides long).
  - Active site of the restriction enzyme is **complementary** to the three-dimensional structure of the specific nucleotide sequence.
  - Most restriction sites are **palindromic**, i.e. **the sequence of nucleotides is the same on both strands read in the 5' → 3' direction**.
- **Hydrolyse phosphodiester bonds** in the sugar-phosphate backbone of both DNA strands at specific points within restriction sites.
- Some restriction enzymes produce **sticky ends** while others produce **blunt ends**.




- Fragments with sticky ends have at least one single–stranded end because the two DNA strands are cut in a staggered manner. The single-stranded DNA can form **hydrogen bonds with complementary sticky ends on any other DNA molecules cut with the same enzyme**. E.g. *Eco*RI, *Bam*HI and *Hind*III restriction enzymes cut DNA into fragments with sticky ends.
- Fragments with blunt ends are produced when the two DNA strands are cut at the same specific position within the restriction site.

(ii) **DNA Ligase**

- Catalyses formation of **phosphodiester bonds** between two nucleotides.
- Used after fragments cleaved using the same restriction enzyme are incubated together to produce a **stable recombinant DNA molecule**
- **Recombinant DNA** refers to manipulated **DNA molecules containing DNA from two or more sources**
- **Complementary sticky ends form temporary hydrogen bonds after which DNA ligase is added to form a permanent phosphodiester bond**.



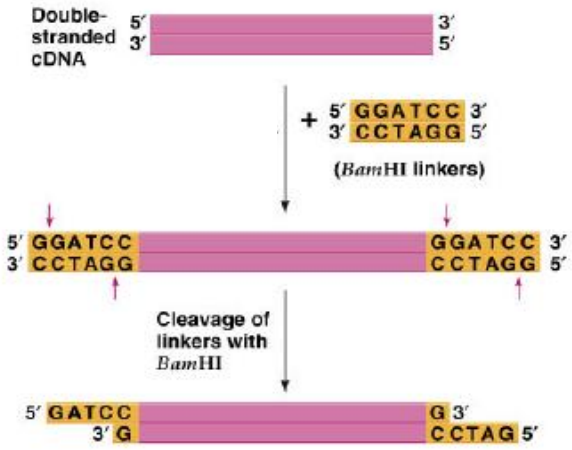

**Formation of recombinant DNA molecule using restriction enzyme and DNA ligase**

## (b) Addition of sticky ends to DNA molecules with blunt ends

In genetic engineering, when restriction fragments have blunt ends, **additional steps are required to form complementary sticky ends**. E.g. *Hae*III and *Alu*I restriction enzymes cut DNA into fragments with blunt ends.

### (i) Addition of specific linker DNA

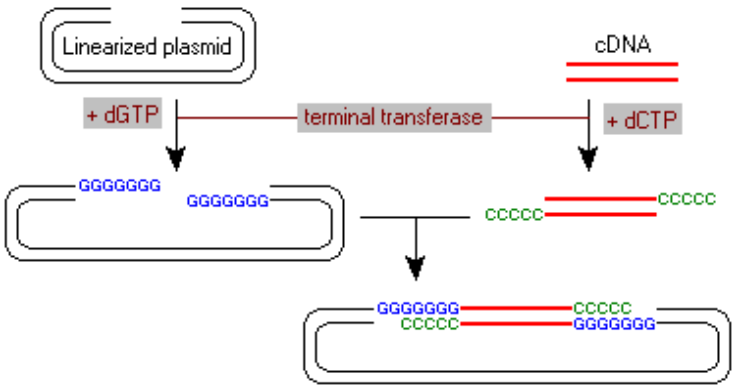
- **Specific linker DNA** is a short, synthetic **double-stranded oligonucleotide** which contains a specific restriction site of a particular restriction enzyme.
- Added to both the blunt ends of the DNA fragment using **DNA ligase**
- Linker DNA is then **further cut by appropriate restriction enzyme** to produce sticky ends.




**Addition of specific linker DNA to form sticky ends**

### (ii) Use of terminal transferase

- **Terminal transferases** are enzymes which catalyse **addition of deoxyribonucleotides** (more specifically deoxyribonucleoside triphosphates) at the **3' ends** of a DNA molecule without using a template.
- Deoxycytidine triphosphate (dCTP) can be added to one blunt-ended DNA fragment and deoxyguanosine triphosphate (dGTP) can then be added to the other DNA fragment to produce **complementary sticky ends**.



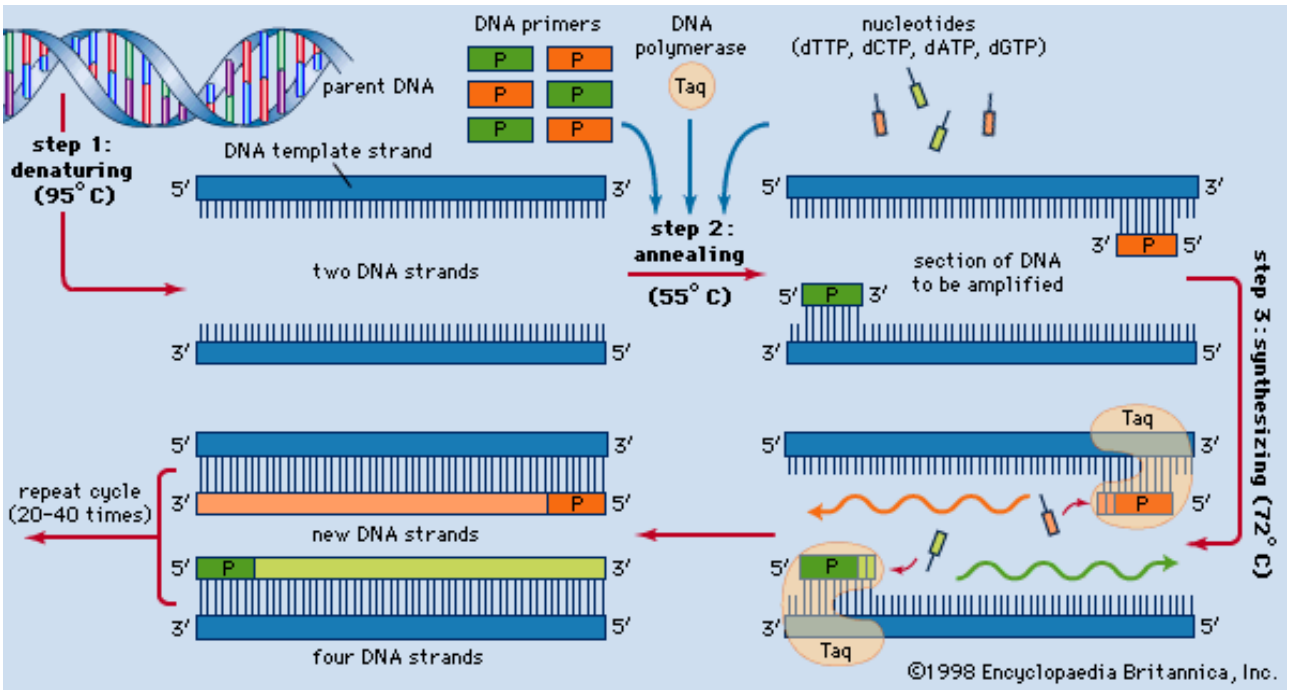

**Addition of deoxynucleotides by terminal transferase to form sticky ends**

## (c) Polymerase Chain Reaction (PCR)

Polymerase Chain Reaction (PCR) is a process used to **amplify a specific segment of DNA** *in vitro*. It results in **synthesis of large amounts of DNA from a minute amount of starting material**. This allows for applications such as basic research and forensics.

### (i) Components of Reaction Mixture

- DNA polymerase such as ***Taq* polymerase** is used. It is a **thermostable** DNA polymerase isolated from the bacterium, *Thermus aquaticus* that thrives in environments of high temperature.
- **DNA primers** are **single-stranded** DNA molecules which are typically 20 to 30 nucleotides long. **Two different DNA primers complementary to the flanking sequence of target DNA sequence to be amplified** are needed.
- **Deoxyribonucleotides** (more specifically deoxyribonucleoside triphosphates) such as dATP, dTTP, dCTP and dGTP are substrates for DNA replication in PCR.




**Process of PCR**

*(Source: http://media-2.web.britannica.com/eb-media/77/22477-004 6584D336.gif)*

(ii) **Process of PCR**

There are three steps in a cycle whereby the reaction mixture will be heated to different temperatures for different periods of time.

1. **Denaturation**
   - The first step involves heating it to a high temperature of about **95°C**.
   - Denaturation of double-stranded DNA into **single-stranded DNA** by **breaking the hydrogen bonds** between the nitrogenous bases of the two strands.

2. **Annealing**
   - Temperature is lowered to about **50 - 60°C**.
   - **Primers anneal via hydrogen bonding,** to the flanking sequence of the target DNA sequence. It binds due to **complementary base pairing** to the single-stranded DNA.
   - Primers are usually used in a much higher concentration as compared to DNA sample.

3. **Elongation**
   - After annealing of primers, temperature is increased to about **72°C** for chain elongation.
   - **Taq polymerase adds nucleotides to the 3'OH end of the primers** using the DNA molecule as a template.

The sequential process of denaturation-annealing-elongation is repeated many times. This is called a chain reaction as the **products of the previous reaction are used as reactants in the next cycle**.

The amount of target DNA sequence doubles after each cycle. Therefore *n* cycles will produce $2^n$ molecules of target DNA sequence. In this way, the amount of target DNA sequence will **increase exponentially**. After about 20 to 30 cycles, the amount of the target DNA sequence will have increased by a billion-fold.

(iii) **Advantage of PCR**

- **Sensitivity**
  - **Large amounts of DNA** can be **produced from very minute amount** of starting materials.
- **Speed and Accuracy**
  - **Large amounts of DNA** can be **produced in a short period of time** with **relatively high accuracy of replication.**
- **Specificity**
  - **Specific sequences of DNA can be amplified** by using **specific primers.**

(iv) **Limitations of PCR**

| Feature | Limitation |
|---|---|
| Use and synthesis of **specific DNA primers complementary to the flanking sequence of target DNA sequence to be amplified** | **Knowledge of the DNA or amino acid sequence** of target gene or protein is **required** to synthesise flanking nucleotide primers.<br><br>If flanking sequences are not known, no amplification will take place. |
| Primers are **short nucleotide sequences** of 15-20 base pairs.<br><br>There may be non-target DNA sequences which are complementary to the DNA primers. | **Non-target DNA sequence may be amplified** alongside the target DNA sequence. |
| **Taq polymerase**, of bacteria origin, **does not perform proofreading**. | There may be **mistakes in complementarity of the nucleotides added and these mistakes may be amplified**. |
| **Limits on the length** of target DNA sequence to be amplified | Target DNA sequence to be amplified by PCR is limited to about 3kb. Efficiency of amplification decreases with increase in length of target DNA sequence. |

## (d) Gel Electrophoresis

Gel electrophoresis is a technique of **separating nucleic acids or proteins based on the size, electrical charge and other physical properties of the molecule** by passing them through a gel which functions as a molecular sieve in an electric field.

It separates macromolecules on the basis of their rate of movement through a gel. **The rate at which a macromolecule travels is inversely proportional to its size**. Larger molecules move at a slower rate while smaller molecules move at a faster rate. Molecules of the same size are separated into a band, in which each band usually contains thousands of molecules of identical size and/or charge.

### (i) Agarose Gel Electrophoresis

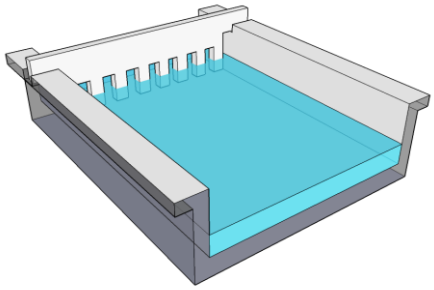
- Agarose is purified from agar, an unbranched polysaccharide present in some species of seaweed. Agarose melts at high temperature and cools to form a gel-like matrix. This porous structure allows it to act as a molecular sieve. This technique is usually used to separate DNA molecules.
- Since DNA is negatively-charged due to its sugar-phosphate backbone, this technique separates the DNA **based on size only** (i.e. number of nucleotides) and not charge.
- The intensity of the band is used to estimate the concentration of DNA sample.

**Process of Agarose Gel Electrophoresis:**

Prior to gel electrophoresis, DNA samples are purified and cut with restriction enzymes.

**(Casting of Gel)**

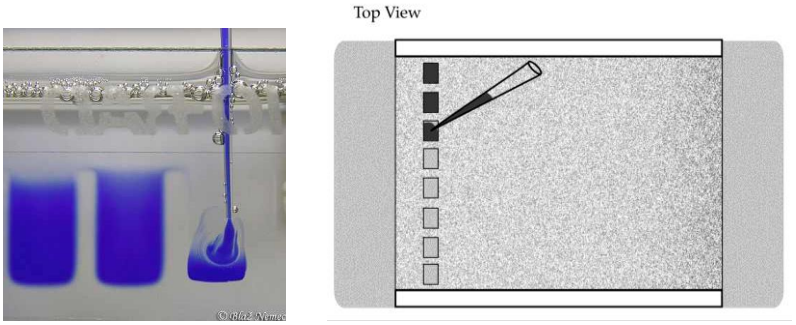
1. The agarose gel is prepared by heating agar powder with buffer solution to dissolve it. The concentration of the gel can be adjusted to alter the resolution of the gel (more agarose will create a less porous gel).
2. The agarose gel solution is then poured into a gel tray and cooled. A comb is added at one end of the gel to create the wells for loading of DNA.
3. After the agarose has cooled and hardened, the comb is removed to reveal the wells. The gel is then placed within an electrophoresis chamber filled with buffer solution.



**Gel Tray with comb at one end of gel**

**(Mixing DNA sample with loading dye and loading DNA into wells)**

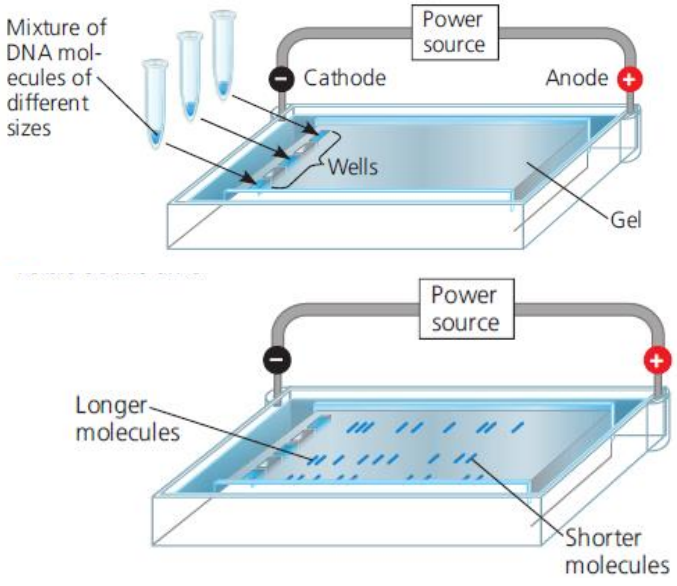
4. DNA samples are then loaded into the wells. The gel is oriented such that the **wells** are placed at the **negative electrode**.
   - As DNA is invisible to the naked eye, a small amount of loading dye, e.g. methylene blue, is usually mixed with the sample. The blue dye will move along the gel and give an indication of the progress of electrophoresis.
   - Glycerol is usually added to loading dye as it is dense and allows the DNA to sink into the wells.
   - In most cases, a standard DNA ladder / kb ladder is also loaded so that the size of DNA fragments can be estimated.




**Loading of DNA into wells**

**(Application of electric field)**

5. The **direct current** (DC) is turned on. After the DNA has run to about two-thirds of the gel length, as visualised by the position of loading dye, the current is stopped.
   - This is to prevent 'over-run' such that DNA samples move off the gel into the buffer solution.




**DNA runs from cathode to anode end when electric field is applied**

*(Bands are shown here but in actual gel, the bands are not visible at this point)*

**(Addition of ethidium bromide for visibility of DNA bands)**

6. After separated in the gel, analysis is not possible as the bands are not visible. Thus, the DNA fragments must undergo a visualisation step.
   - **Ethidium bromide** can be added in the buffer or added in the agarose gel before cooling. When viewed under **UV radiation**, DNA bands **fluoresce**. A polaroid photo can then be taken to record the position of the bands.

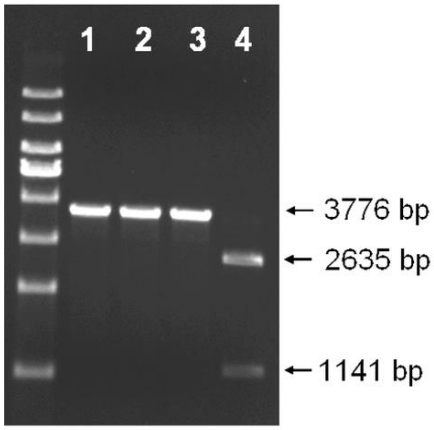
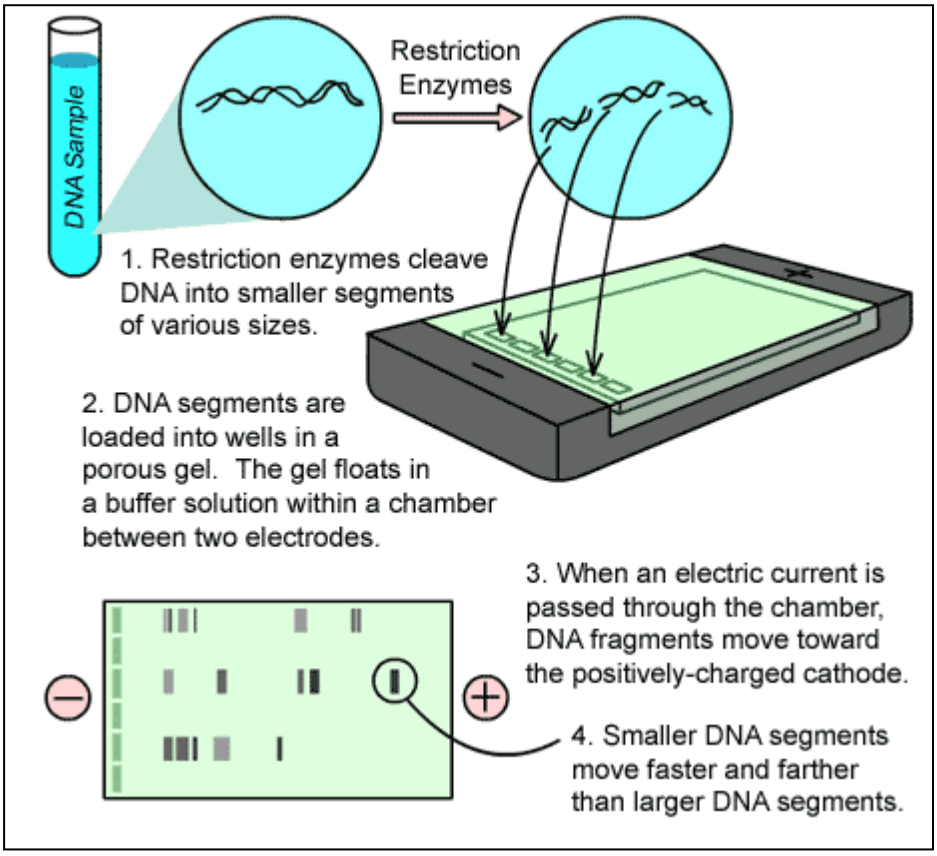



**Photo of DNA bands made visible with ethidium bromide**

*(Bands on extreme left lane came from DNA ladder)*

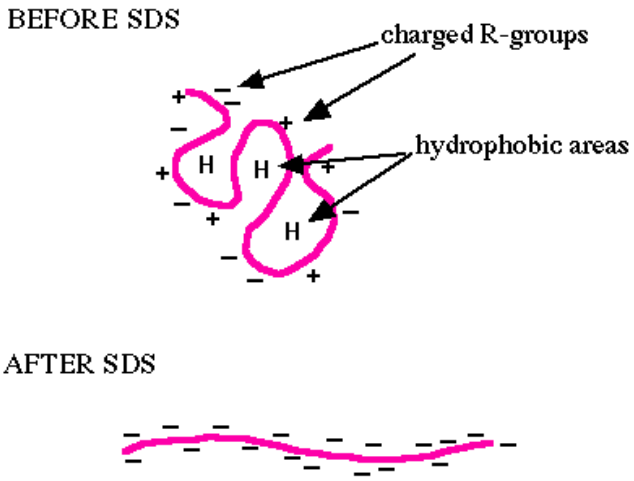



**Overview of Agarose Gel Electrophoresis**

*(http://www.stanford.edu/group/hopes/diagnsis/gentest/f_s02gelelect.gif)*

(ii) **SDS PAGE**

- Sodium Dodecyl Sulphate Polyacrylamide Gel Electrophoresis (SDS PAGE) is usually used to separate proteins of different sizes.
- Polyacrylamide is an organic polymer. Gels made from polyacrylamide have a higher resolving power as compared to agarose gels.
- Sodium dodecyl sulphate (SDS) acts as a denaturing chemical that breaks bonds holding the quaternary, tertiary and secondary structures of proteins.
- SDS is negatively charged, and it causes the unfolded polypeptide chain to be negatively charged. In this way, the polypeptide chain can move towards the positive electrode during gel electrophoresis. The above allows the separation of proteins based on size only (i.e. the number of amino acids in the primary structure) and not charge.



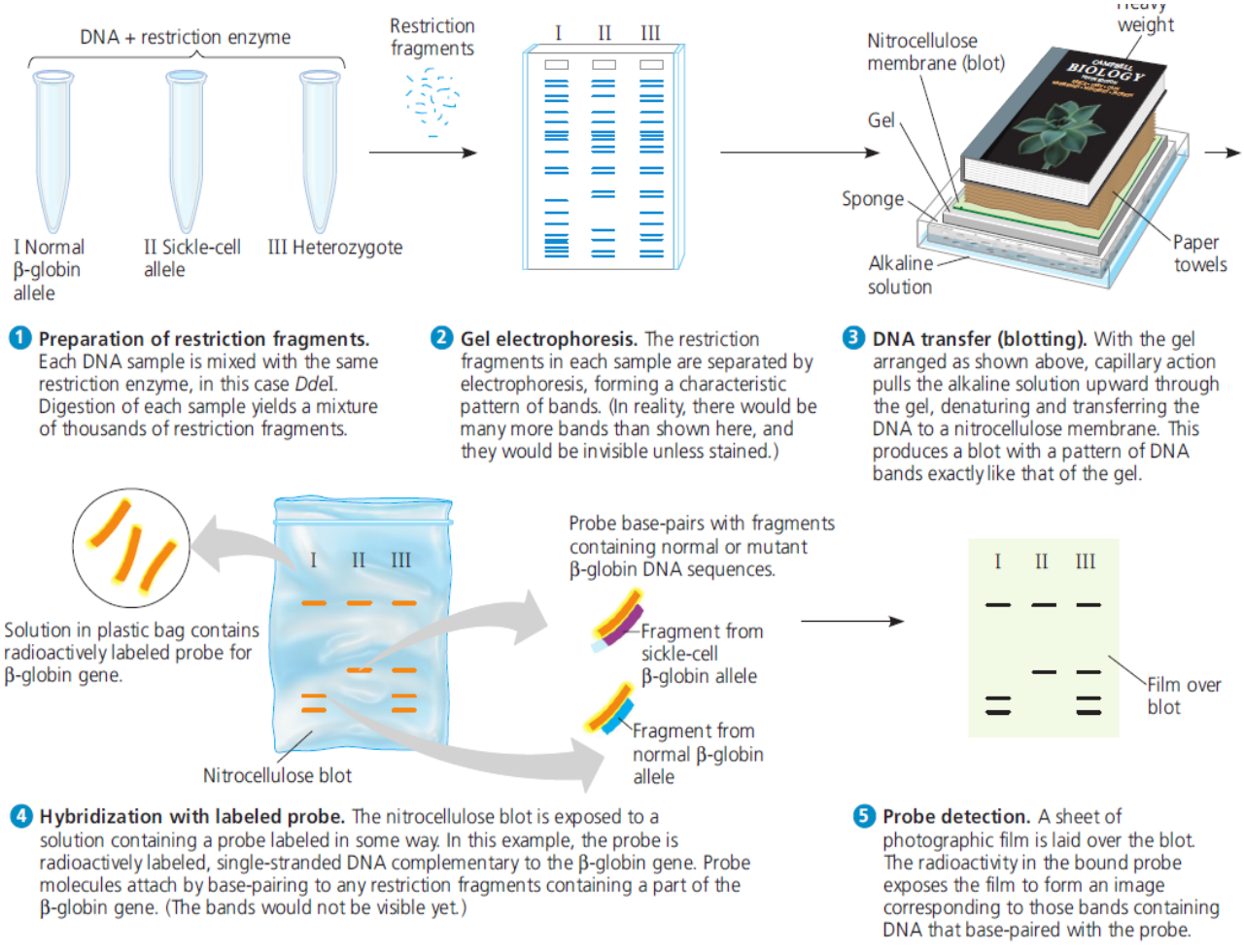

**Negatively charged proteins after SDS treatment**

## (e) Nucleic Acid Hybridisation

**Nucleic acid hydridisation** is the **complementary base pairing** of a nucleic acid to another nucleic acid to form a double stranded molecule. Nucleic acid hybridisation is carried out using a **nucleic acid probe** which is a short, single stranded nucleic acid that can be ether DNA or RNA to locate the gene of interest.

**Southern blotting** is a method that combines gel electrophoresis and nucleic acid hybridisation. It allows the detection of DNA bands which contain the gene of interest. Southern blotting involves **transfer of separated DNA fragments onto a membrane for hybridisation**.

Nucleic acid hybridisation in Southern blotting is carried out using a DNA probe. **A DNA probe is a single-stranded DNA sequence which is complementary to the gene of interest and is usually labeled**.




**Process of Southern Blotting**

**Process of Southern Blotting with Nucleic Acid Hybridisation:**

(Steps correspond with diagram on previous page)

**1. Preparation of restriction fragments**

- DNA samples are purified and cut with restriction enzymes.

**2. Gel electrophoresis**

- DNA fragments are separated using gel electrophoresis.

3. **DNA transfer from gel to nitrocellulose membrane**.

- **Double-stranded DNA** is **denatured into single strands** with addition of **sodium hydroxide**. Hydrogen bonds between complementary bases are broken.
- **Bands on the gel are transferred to a nitrocellulose membrane** via capillary action or using vacuum suction.
  - The gel is placed in a basin of buffer solution containing sodium hydroxide. A piece of **positively charged nitrocellulose membrane** is placed directly above the gel.
  - Large amounts of dry paper towels and heavy weights are then placed on top of the setup to generate **capillary action**. The dry paper towels absorb buffer solution through the gel and membrane.
  - Capillary action will transfer the single-stranded DNA bands onto the nitrocellulose membrane **corresponding to those on the gel**.
- The negatively charged DNA forms temporary bonds with the positively charged membrane. The nitrocellulose membrane is then **subjected to high temperature or UV light** to **permanently cross-link single-stranded DNA to the membrane**.

4. **Nucleic acid hybridisation**

- **Nitrocellulose membrane** containing the **single-stranded** DNA is **incubated** in a solution containing the **labeled DNA probe**
- Depending on what the probe is labeled with, different methods of detection is used. In this case, DNA probe is **labeled with nucleotides containing radioactive $^{32}$P**.
- The probe **binds to the complementary DNA fragment of interest via hydrogen bonds**.
- Unhybridised probes are washed away.

5. **Probe detection**

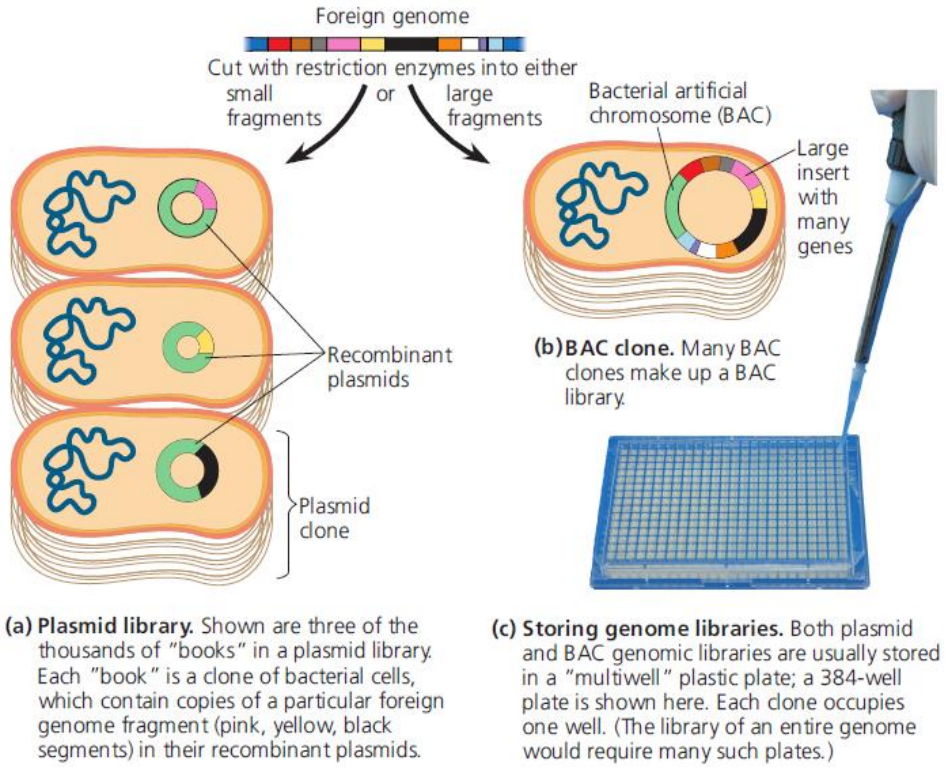
- When exposed to a piece of **X ray film**, the radioactive isotope exposes the film. The band containing the gene of interest will show up as a dark band on the autoradiogram.
- Colorimetric methods involve a DNA probe **labeled with luminescence molecule**. This luminescence molecule emits light and the band can be captured on Polaroid photographs or detected using laser cameras.

## 3. DNA Libraries

### (a) Genomic DNA Library

The **complete set of genetic material** inclusive of **all coding and non-coding** (repetitive sequences, control elements, introns) sequences is known as **genomic DNA**.

- The **genomic library** consists of a set of clones containing **DNA fragments representing the entire genome of an organism**.
- It is generated by subjecting genomic DNA to physical shearing or **cleavage by restriction enzymes to generate DNA fragments**.
- The DNA fragments are then inserted into **vectors** such as **plasmids**. Other vectors like **bacterial artificial chromosome (BAC)** and **yeast artificial chromosome (YAC)** which can contain more base pairs, about 350-1000kbps, are also used. Another alternative vector is the **λ phage**. However, only relatively short base pairs, about 10-20kbps, can be inserted.
- Genomic libraries are useful when:
  - **Source of gene of interest is not known**. When it is not known which cell type expresses the gene of interest, a genomic library is used as it represents the entire genome.
  - Purpose of research is to **study the regulatory sequences or introns** of a gene

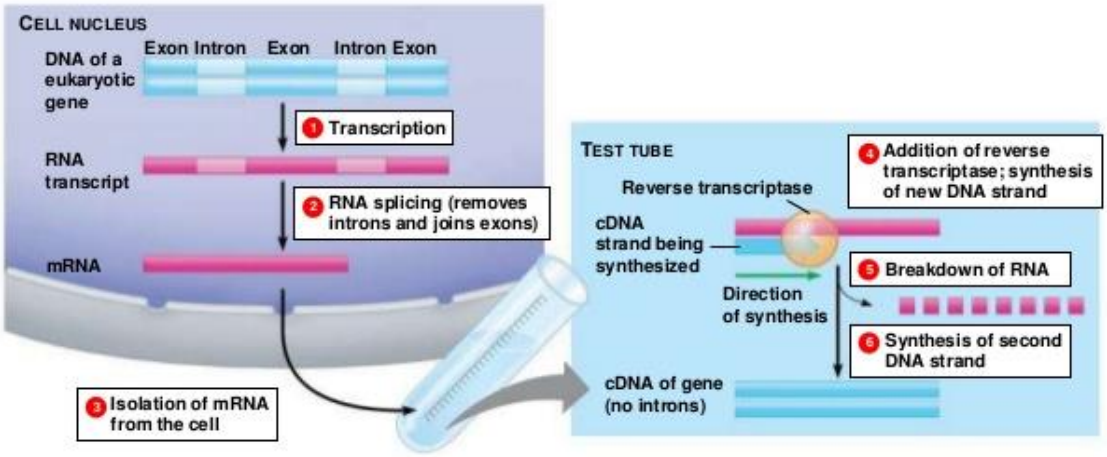



**Genomic libraries**

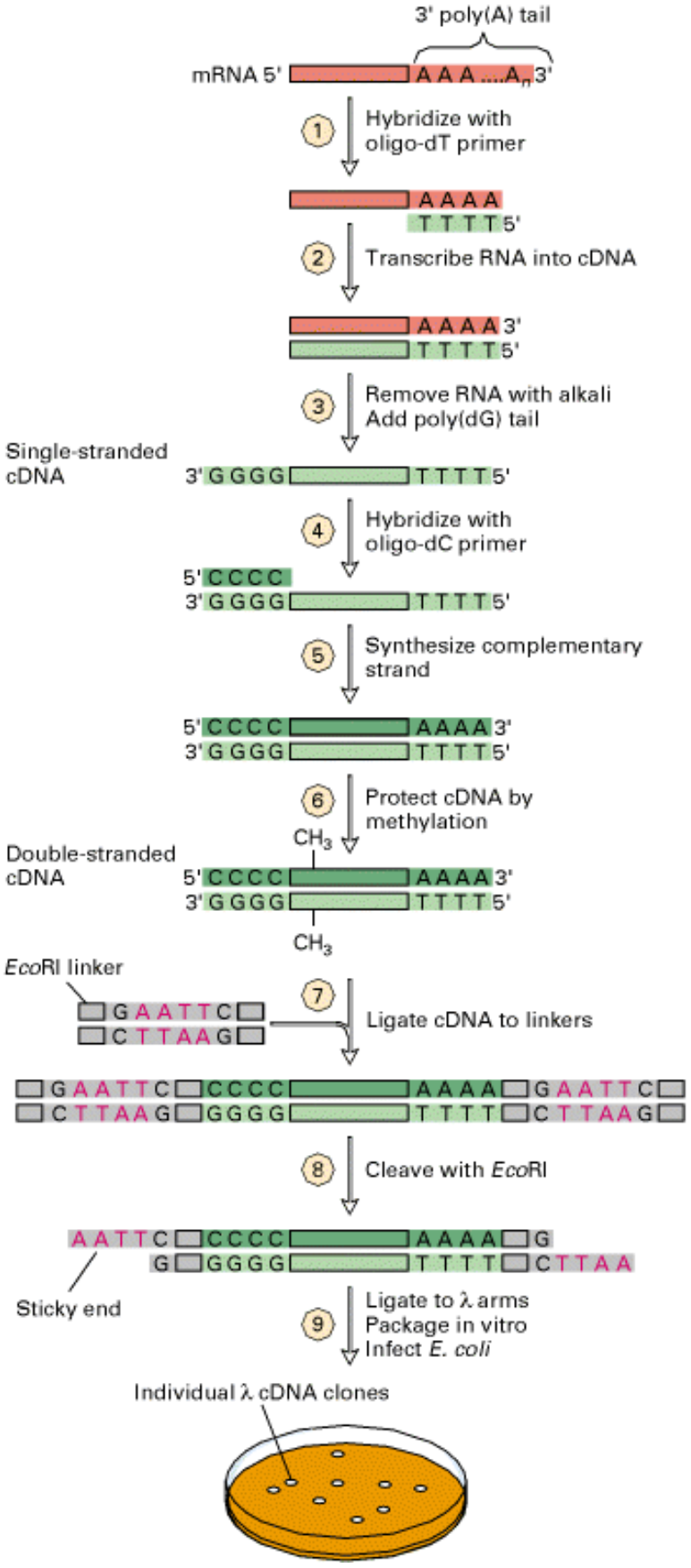
(b) **Complementary DNA Library (cDNA library)**

The **complementary DNA library** is a gene library containing clones that carry only **complementary DNA molecules synthesized from mRNA molecules in a cell**.

- The construction of a **cDNA library** involves **reverse transcriptase** which is used to convert mature mRNA to DNA called complementary DNA (cDNA).
- cDNA is then inserted into **vectors** to be stored for later use. As cDNAs are relatively short, **plasmids** and **phages** are usually used.
- Mature messenger RNA (i.e. without introns) can be obtained from any particular cell. Most cells will produce essential proteins (e.g. tubulins in cytoskeleton).
- cDNA libraries are useful when:
  - **Isolating a set of genes that are responsible for specialised function of a particular cell type**. E.g. tissue-specific mRNAs can be obtained from certain cells (e.g. mRNA of insulin receptor in liver cells, mRNA of pepsinogen in stomach cells).
  - **Tracing changes in patterns of cell expression during development** by looking at cDNA libraries of the **same cell type at different times in the life of an organism**




**Making cDNA from mRNA**



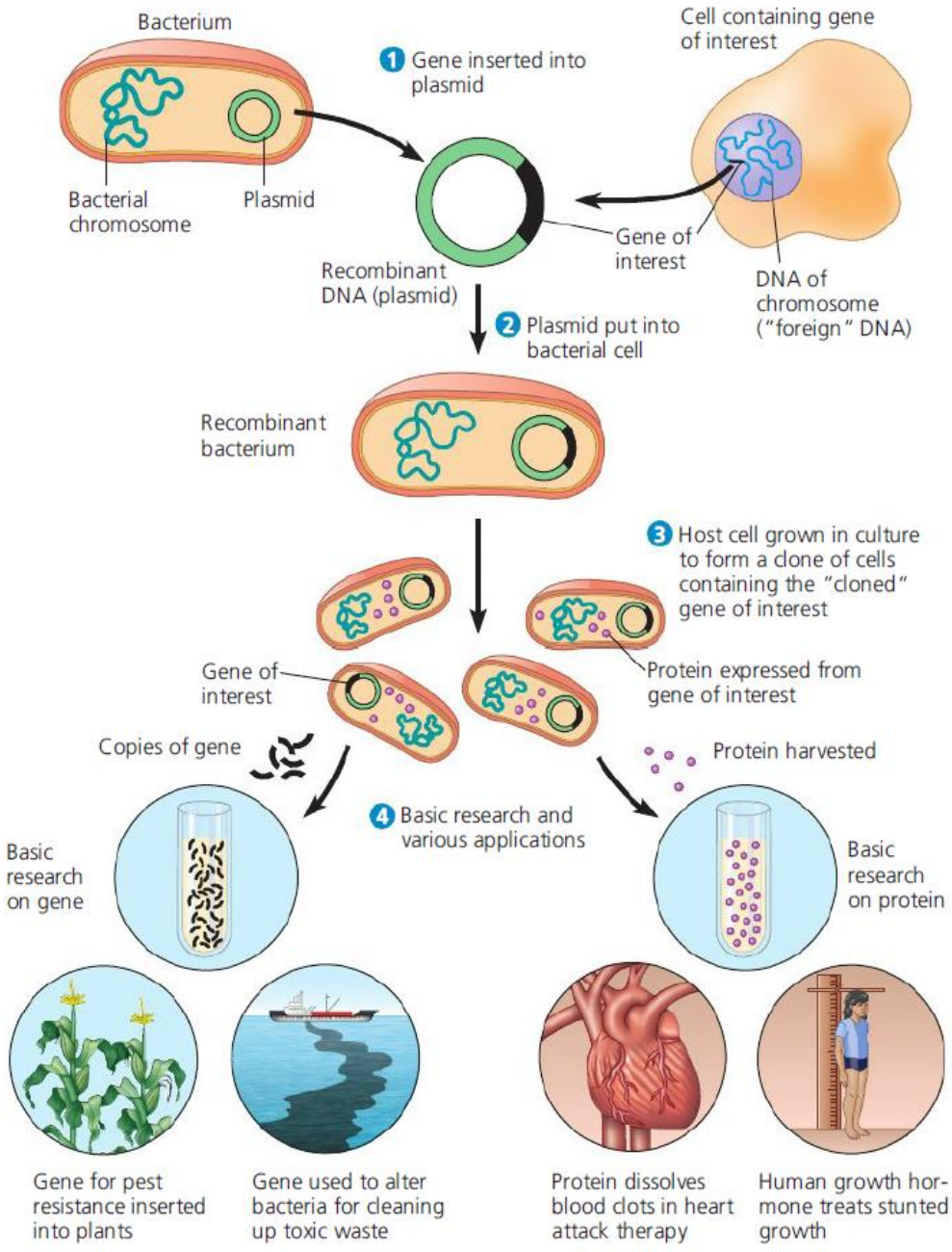

**Construction of cDNA library using the λ phage as vector**

## 4. Gene Cloning

To work directly with specific genes during genetic engineering, scientists have developed methods such as recombinant DNA technology for production of multiple identical copies of whole genes, a process called **gene cloning**.

Gene cloning is useful for two basic purposes, which is to **make many copies of a single gene** or to **produce large quantities of protein product coded by the gene**.

This section discusses the use of plasmids and the steps taken to clone a eukaryotic gene into a prokaryotic bacterial cell.




**Gene cloning and uses of cloned genes**

(a) **Plasmids**

Bacterial plasmids are naturally occurring **extra-chromosomal, small, circular** DNA which contains genes which **confer selective advantage** to the bacterial cells. They function as **cloning vectors** which carry foreign DNA into a host cell and replicate within the host cell. Some properties of plasmids make them suitable for use as a cloning vector in genetic engineering.

**Properties of Plasmids**

(i) **Small circular DNA**

- Plasmids can be **induced to enter host cells easily** by transformation.

(ii) Contain **restriction enzyme recognition sites**.

- Allows **cleavage and insertion of foreign genes** using restriction enzymes and DNA ligase

(iii) Contain their **own origin of replication**

- Allows them to **replicate autonomously**, independent of the bacterial chromosome.

(iv) Exists in **high copy number**

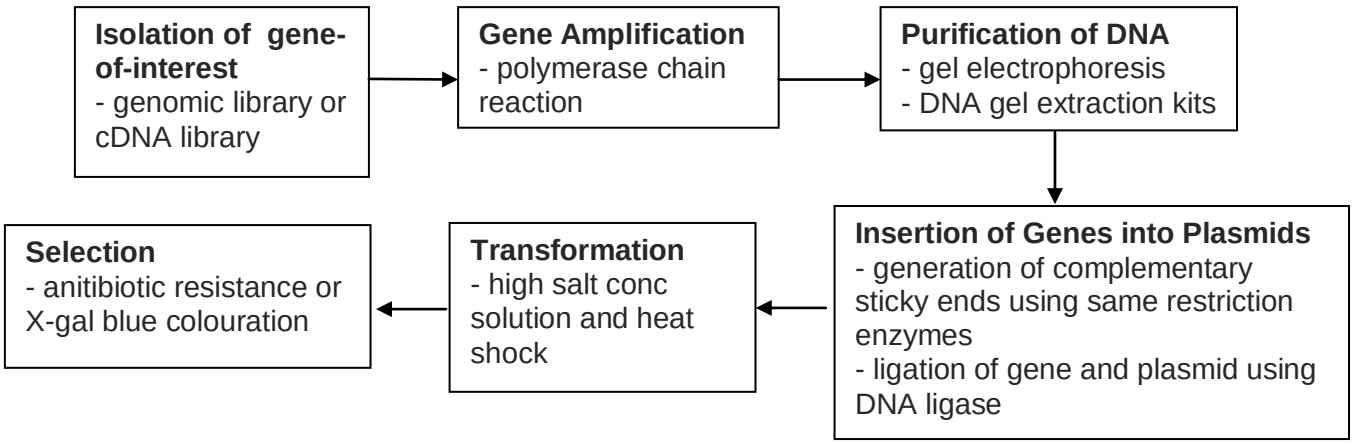
- **Quantity of plasmid DNA that can be purified from each host cell is high**.

(v) Contain one or more **selection marker genes**.

- Allows for production of proteins giving **a phenotype used to identify plasmids which has successfully incorporated the gene of interest** and **cells which had taken up the recombinant plasmid** via transformation.
  - Antibiotic resistance genes conferring resistance to ampicillin and tetracycline may be present in plasmids such as pBR322.
  - Selection marker genes may cause formation of coloured products. E.g. pBlueScript is a plasmid containing the lacZ gene which codes for β-galactosidase. When β-galactosidase is expressed, a colourless lactose-like substrate, X-gal in a medium is cleaved to form a blue product, resulting in blue bacterial colonies, thus allowing identification.

(b) **Process of Gene Cloning**

**Schematic Diagram of Steps in Gene Cloning**

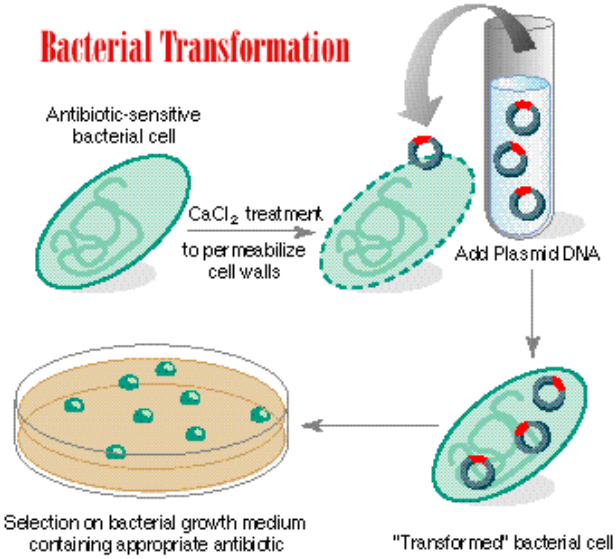



- **Transformation**

  The **process of introducing the recombinant plasmid into a host cell** is known as **transformation**. Host cells induced to take up plasmid DNA from the surrounding medium are known as **competent cells**.

  - Bacteria cells can be made competent by placing them in high salt condition using **calcium chloride solution.** Plasmid DNA is then added in and the mixture is subjected to **heat shock** (e.g. 42°C for 30 – 120 seconds), followed by the lowering of temperature by placing it on **ice**.

  - Positively charged calcium ions neutralize negatively-charged DNA. When bacteria are subjected to heat shock, it **induces temporary gaps in the cell membrane of host cells and allow plasmids to enter**.




*http://www.assessexcellence.org/RC/AB/WYW/cohen/transformation.gif*

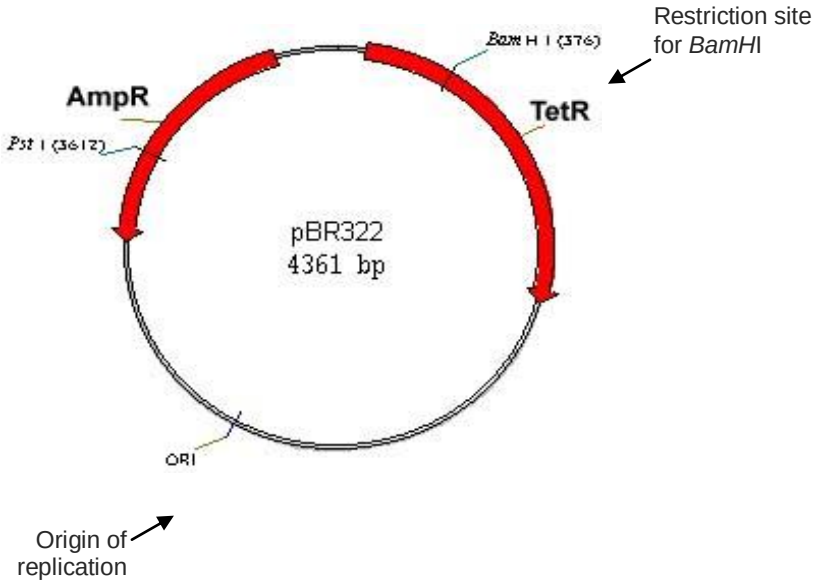
- **Selection**

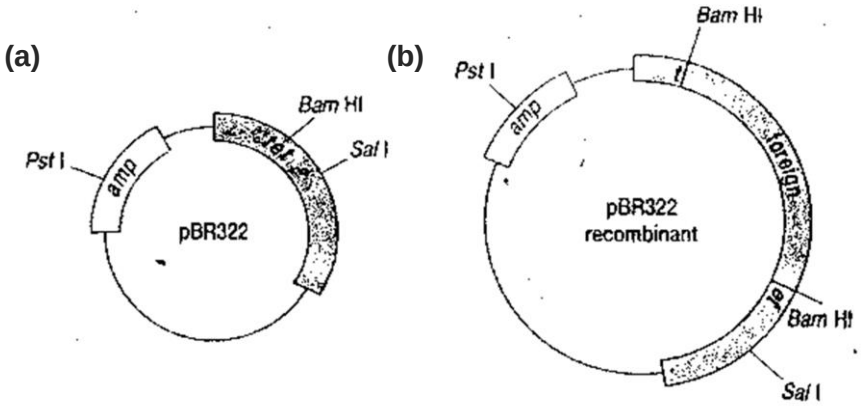
  The process of recombinant plasmid formation and transformation is a **random** process. There are many other non-recombinant plasmids which may be taken up by the host cell. Also, not every competent host cell will take up a plasmid. Thus, there must be some form of selection to ensure **bacterial cells containing recombinant plasmid with gene of interest inserted can be identified**.

**(i) Antibiotic Resistance Selection / Screening**

Replica plating is conducted when two antibiotic resistance genes are used as selection markers.

- Replica plating is a technique in which one or more secondary Petri dish containing different selective growth media, are inoculated with the same colonies from a primary plate, **reproducing the original spatial pattern of colonies**.
- The following plasmid is used as an example. pBR322 contains the ampicillin resistance gene and tetracycline resistance gene




*(a) Non-recombinant pBR322 with intact tetracycline resistance gene*

*(b) Recombinant pBR322 with inserted gene of interest, disrupting tetracycline resistance gene*

- Transformed bacterial cells are inoculated, or placed on a culture medium containing ampicillin where it would reproduce. It is incubated overnight at 37°C.
- Successfully transformed cells containing the plasmid will survive the ampicillin plate (primary plate) and grown into individual bacterial colonies.
- **Ampicillin** in the medium **selects for bacterial cells that are successfully transformed** and contains the plasmid. Hence all colonies that are present possess the plasmid containing the ampicillin resistance gene. Untransformed bacterial cells are susceptible to ampicillin and dies.
- A **sterile velvet cloth** or a piece of **sterile filter paper** is used to make a replica of the ampicillin plate. This replica is then gently pressed onto an agar plate containing tetracycline (secondary plate). This ensures that the bacterial colonies are at the same positions relative to one another.

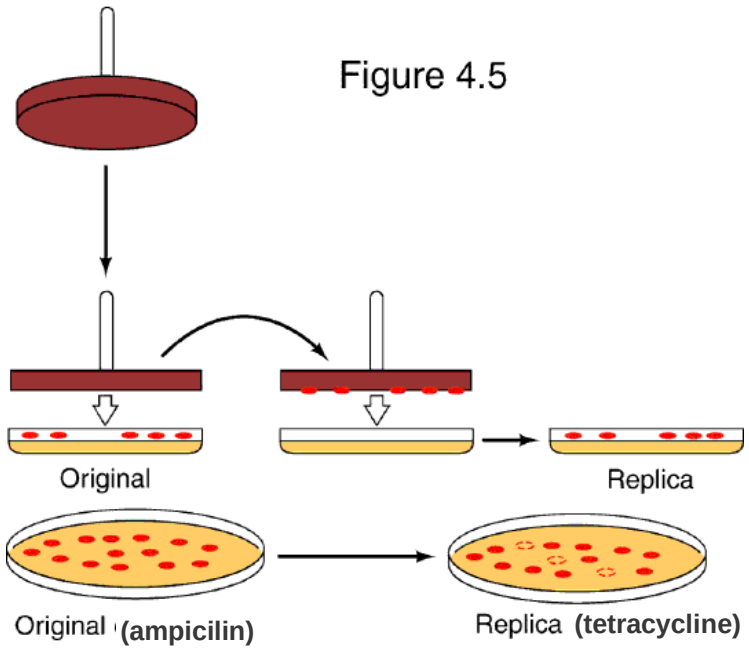
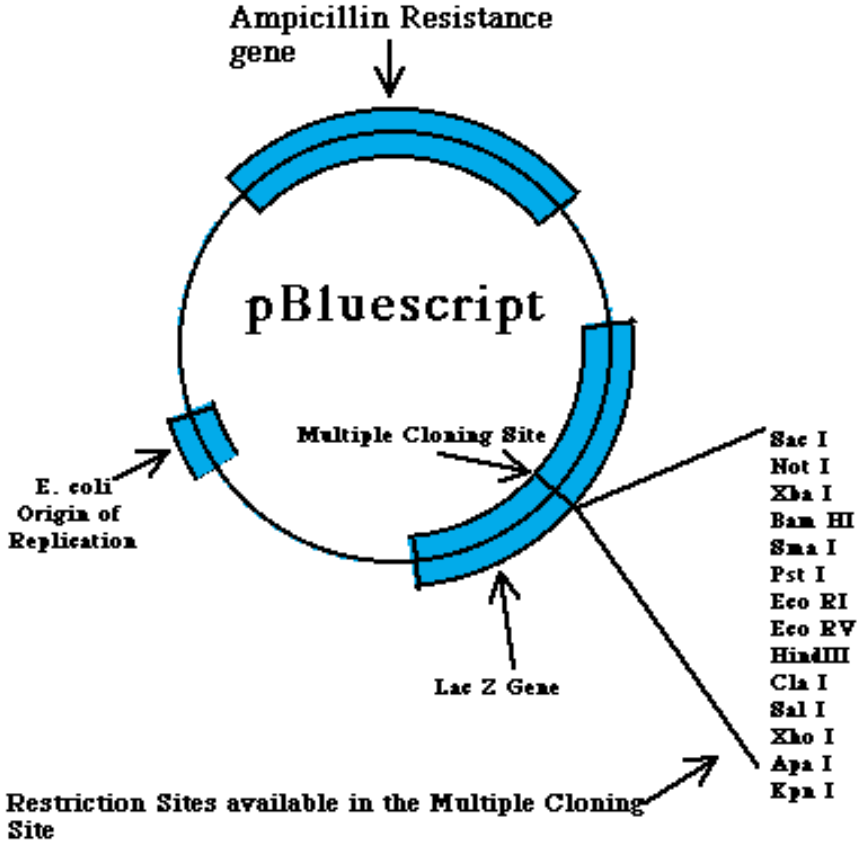


Figure 4.5

- There is a restriction site for *Bam*HI within the tetracycline resistance gene.
- Therefore, when plated on a tetracycline plate,
  - If it is a **reannealed non-recombinant plasmid** with no gene-of-interest is inserted, the **tetracycline resistance gene remains intact** and **functional gene product** is synthesised. Bacterial cell survives on tetracycline plate
  - If it is a **recombinant plasmid** with gene-of-interest is inserted, it will **disrupt the tetracycline resistance gene**. **Insertional inactivation** of gene results in a **non-functional gene product.** Bacterial cell **become susceptible to tetracycline and die**
- If a bacterial clone survives on ampicillin medium but does not survive on tetracycline medium, it shows that the gene-of-interest had been inserted within the tetracycline resistance gene
- Comparison is done to pick out bacterial colonies that are present on the ampicillin plate but absent in the tetracycline plate for further manipulation.

**(ii) Blue-white Selection / Screening**

Commercially available plasmid pBluescript is used as a vector and uses **blue-white screening** to select for successfully transformed recombinant bacterial cells**.** It contains an **ampicillin resistance gene** and **a *lac*Z gene**. A **multiple cloning site** (MCS) is present within the *lac*Z gene. A MCS is a short segment of DNA which contains many restriction sites. The restriction sites within an MCS usually occur once within the plasmid.




**pBluescript plasmid**

*(http://www.bio.davidson.edu/Courses/Molbio/MolStudents/spring2003/Keogh/pBluescript.gif)*

- The bacterial host cell is usually a strain of *E.coli* without the *lac*Z gene and hence do not synthesize any β-galactosidase. The transformed bacterial cells are inoculated on nutrient medium containing IPTG, X-gal and ampicilin.
  - IPTG used to induce synthesis of β-galactosidase.
  - X-gal is a **colourless substrate** which when cleaved by β-galactosidase, produces a **blue insoluble product**.

- **Ampicilin** in the medium **selects for bacterial cells that are successfully transformed** and contains the plasmid. Hence all colonies that are present possess the plasmid containing the ampicillin resistance gene. Untransformed bacterial cells are susceptible to ampicillin and dies.
  - If it is a **reannealed non-recombinant plasmid** with no gene-of-interest is inserted, the **lacZ gene remains intact** and **functional β-galactosidase** are synthesised. X-gal will be cleaved by β-galactosidase and a **blue** colony is observed.
  - If it is a **recombinant plasmid** with gene-of-interest is inserted within the MCS, it will **disrupt the lacZ gene**. **Insertional inactivation** of gene results in a **non-functional β-galactosidase** protein being synthesized. **X-gal is not hydrolysed** and the bacterial colony remains **white**.
- The white colony is then picked and cultured for further study.

## (c) Production of Proteins using Genetic Engineering

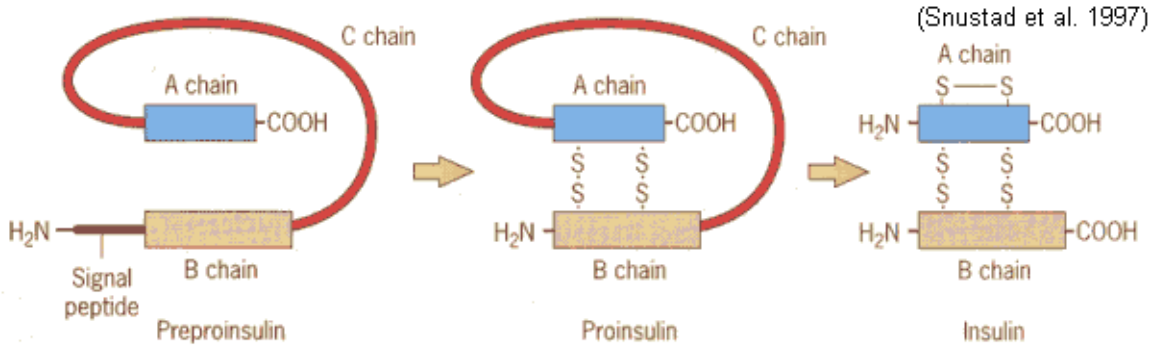
One important use of genetic engineering (gene cloning) is the production of useful protein products. Many medically useful proteins can now be effectively produced using recombinant DNA technology.

Common problems of using prokaryotes to synthesize eukaryotic proteins are as follows:

- **Eukaryotic genes possesses introns** which bacterial cells do not have the cellular machinery for splicing of introns.
- Most eukaryotic proteins consist of various subunits which are covalently attached via post-translational modification. Bacterial cells are **unable to carry out post-translational modifications** as they lack several eukaryotic organelles.
- Prokaryotic translation is initiated by **addition of N-formyl-methionine**, which is the first amino acid added as compared to methionine in eukaryotes, so the protein synthesized may be different.

### (i) Human Insulin

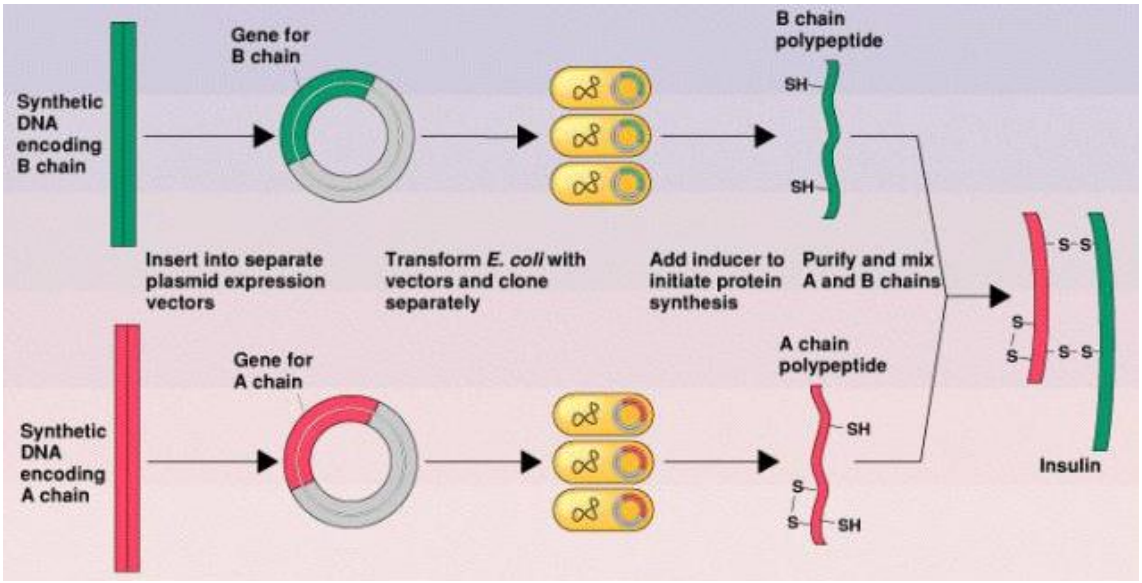
Insulin is a peptide hormone that regulates blood glucose levels. In the human cell, insulin gene has two exons separated by an intron. Primary mRNA transcript undergoes post-transcriptional modification (splicing) to form the mature mRNA. When translated, preproinsulin is formed. Preproinsulin undergoes a series of post-translational modification to form insulin.




**Post-translational modifications to preproinsulin to form insulin**
*(http://www.mun.ca/biology/scarr/Insulin_posttranslational_modification.htm)*

In the rER, the signal peptide of preproinsulin cleaved to form proinsulin which contains A chain and B chain, separated by C chain. Proinsulin is then transported to the Golgi Apparatus where the A chain and B chain are joined via disulfide bonds and the C chain removed.

For the formation of synthetic insulin, known as humulin, it is not possible to use the genomic human insulin gene as source of DNA due to the need for complex post-transcriptional and post-translational steps.




**Production of insulin from genetic engineering**

*(http://courses.cm.utexas.edu/emarcotte/ch339k/fall2005/Lecture-Ch8-2/Slide8.JPG)*

- Using the amino acid sequence of **A chain** and **B chain**, the **DNA sequence of the polypeptide chains are artificially synthesised**. **A strong bacterial promoter** is added at the 5' end for each of the DNA sequence of A and B chain.

- The DNA fragments for the A and B chain are **amplified by polymerase chain reaction**.

- These two DNA fragments are **cloned separately** into two bacterial plasmids.
  - To form each of the recombinant plasmid, **specific linker DNA** is added to blunt ends of DNA fragment. **DNA fragment and plasmids are then mixed together** and cleaved with the **same restriction enzyme** to produce **complementary sticky ends**. **DNA ligase** is added to seal the DNA fragment and plasmid.

- The plasmids containing the DNA sequence of the A or B chains are introduced into bacterial host, ***E. coli*** via **transformation**. Bacterial colonies containing **successfully transformed plasmids are selected** using different methods discussed earlier, depending on the selection markers present.

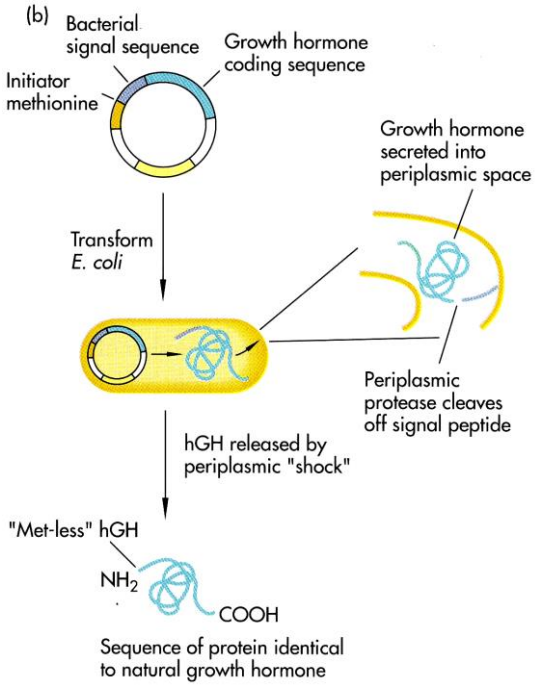
- These **bacterial cells are cultured** and the fusion bacterial-human polypeptide is synthesized by the bacterial cells. The **polypeptide is harvested** from bacterial cells **by lysing the bacterial cells**. A and B chains are then **purified** from the lysate.

- **Bacterial amino acids and N-formyl methionione are removed**.

- The **purified chains are mixed together in equal portions** and **incubated under suitable conditions for formation of disulfide bonds** to form synthetic insulin.

(ii) **Human Growth Hormone**

**Human growth hormone** is a peptide hormone secreted by the pituitary gland which regulates the normal growth of an individual. Deficiency in this factor results in dwarfism, however, it can be corrected by administering this hormone to affected individuals.




**Production of human growth hormone from genetic engineering**

- Large amounts of human growth hormone **mRNA is extracted from the pituitary gland**. Reverse transcriptase is then used to produce **cDNA**.

- **Specific linker** DNA is added to blunt ends of cDNA. **cDNA and plasmids are mixed together** and cleaved with the **same restriction enzyme** to produce **complementary sticky ends**. The cDNA is then joined to the plasmid using **DNA ligase**, at a **site after a bacterial promoter and signal peptide sequence**.

- The recombinant plasmids are introduced into bacterial host, ***E. coli*** via **transformation**. Bacterial colonies containing **successfully transformed plasmids are selected** using different methods discussed earlier, depending on the selection markers present.

- After translation, the **protein is transported to the periplasmic space of bacterial cell due to the signal peptide**. The **signal peptide is cleaved** at the periplasmic space.

- The cell is placed into a **hypotonic solution**. This **disrupts the outer membrane** and releases the protein into the solution.

- **Further purification** is carried out to obtain pure and sterile hormone.

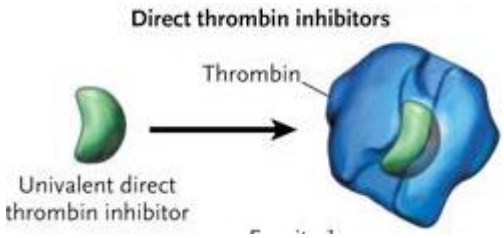
(d) **Other Products of Genetic Engineering**

**Anti-thrombin**

Thrombin is an enzyme that cleaves fibrinogen to form fibrin. It is involved in activating blood coagulation. It also activates other coagulation factors like platelets. Anti-thrombin is a small glycoprotein produced in the liver which inhibits thrombin by association with herapin (an anticoagulant), and by binding to the active or allosteric sites of thrombin. As such, anti-thrombin prevents coagulation/ clotting of blood.

Genetically engineered female goats with anti-thrombin DNA cloned into their genes produce human anti-thrombin in its milk. The cDNA of anti-thrombin, with a promoter which directs secretion of proteins into milk is used. An example of such a promoter is that of casein.

The human anti-thrombin is then purified from the goats' milk to manufacture ATryn, the anti-thrombin drug manufactured by GTC Biotherapeutics. FDA recently approved this drug in early 2009. It is used to treat hereditary anti-thrombin deficiency, deep vein thrombosis (DVT) and reduce risk of clotting during childbirth and surgery.




Adapted from: http://content.nejm.org/content/vol353/issue10/images/large/09f2.jpeg

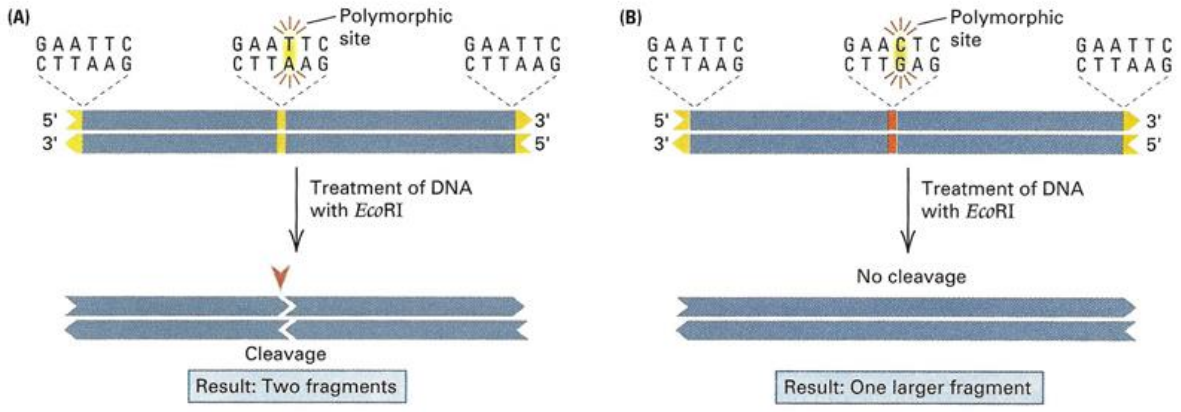
## 5. Restriction Fragment Length Polymorphism Analysis

DNA polymorphisms are variations in DNA sequence among a population. On average, variation frequency is about 1 in every 1000 base pairs in the human genome.

Restriction fragment length polymorphism (RFLP) is a **variation in the DNA sequence of a genome** that can be **detected by cleaving DNA molecules with restriction enzymes** and **analysing the size of the resulting fragments by gel electrophoresis**.

It arises due to **differences in nucleotide sequences at restriction sites** on homologous chromosomes of individuals. A **restriction enzyme may be altered** such that it is **no longer recognised by its restriction enzyme**.

This means that when the chromosome (with restriction site containing DNA polymorphisms) is cleaved with the corresponding restriction enzyme, the **restriction fragment pattern** is **unique** and differs between homologous chromosomes of an individual **and** between individuals such that the **number and position of the bands present may differ**.




**RFLP as a result of DNA polymorphisms occurring within restriction site**

The analysis of the unique restriction fragment patterns is known as RFLP analysis. This technique allows for applications in disease detection, DNA fingerprinting and genomic mapping.

To investigate RFLP between individuals, the process generally involves:

- Cleavage of DNA molecules with suitable restriction enzyme
- Southern Blotting:
  - Separation of restriction fragments using gel electrophoresis
  - Nucleic hybridisation to detect the bands containing the sequence of interest using a DNA probe complementary to the sequence of interest

Polymorphism arises due to various reasons

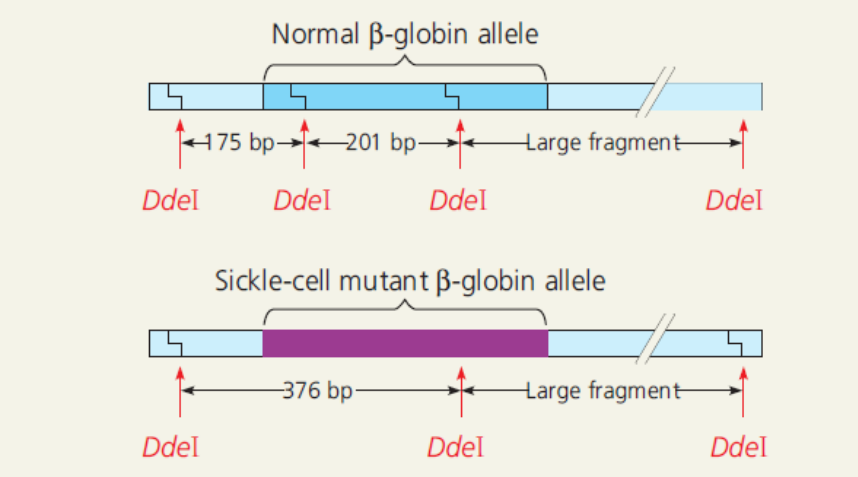
- Variation in nucleotide sequences of genes and non-coding regions of individuals may occur through **point mutations**. A **restriction site may be altered** such that it is **no longer recognised by its restriction enzyme** and therefore **changes the banding pattern of the RFLP**. (Elaborated in the previous page)
- **Deletions** and **duplications** may delete or add segments of DNA to a particular region of the chromosome. **Shorter or longer RFLPs changes the position of the bands**.
- **Intergenic sequences** consist of tandem repeats of specific sequences. These sequences are known as **variable number of tandem repeats (VNTR)** because the **number of repeats between individuals differs**. The analysis of differences arising from differing repeat number is known as VNTR analysis and is a specialized form of RFLP.

(a) **Disease Detection**

RFLP analysis can be used to distinguish between alleles of a particular gene. In order to make use of RFLP for disease-causing gene detection, the polymorphism (difference in nucleotides) should either be **within or closely linked to alleles that are related to a particular disease**. This technique thus helps to identify the alleles that cause diseases.

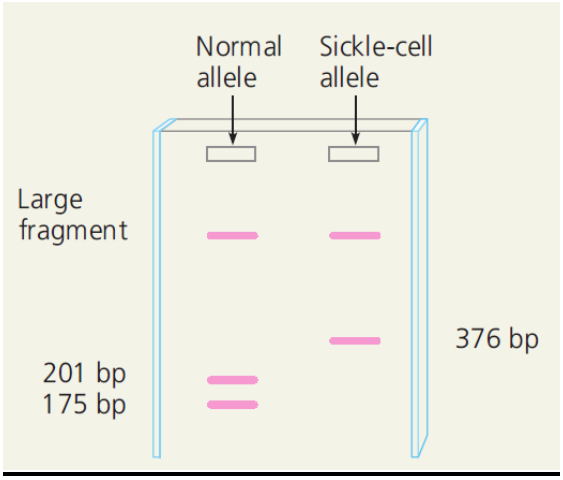
(i) **Sickle-cell Anaemia** (RFLP located within allele)

- Disease caused by point mutation of a single base pair in the β-globin gene
- Mutated base pair happens to **lie within the restriction site of a restriction enzyme, *Dde*I**
- **Normal β-globin allele has four restriction sites** for *Dde*I
  - **Produces 3 fragments** upon digestion: 2 shorter ones (291bp and 175bp) and a much longer one
- **Mutant β-globin allele only has three restriction sites** for *Dde*I
  - **Produces 2 fragments** upon digestion: 1 shorter fragment (376bp) and a much longer one




***Dde*I restriction sites in normal and mutant β-globin allele**

- When analysed by gel electrophoresis, the **band patterns (number of bands and position of bands) formed by normal allele and sickle cell allele are different**. Thus, this method can be used to identify individuals carrying the sickle cell mutant β-globin allele.

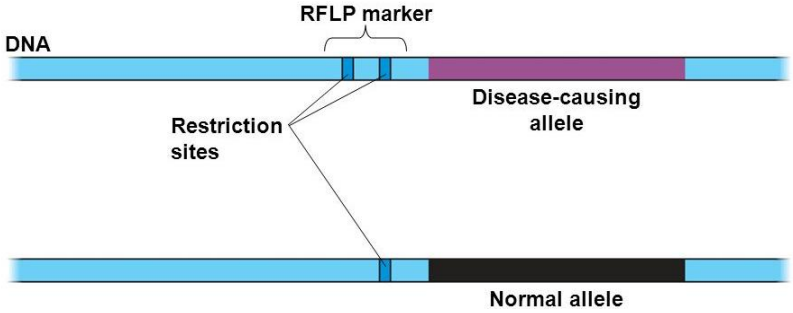



***Dde*I restriction sites in normal and mutant β-globin allele**

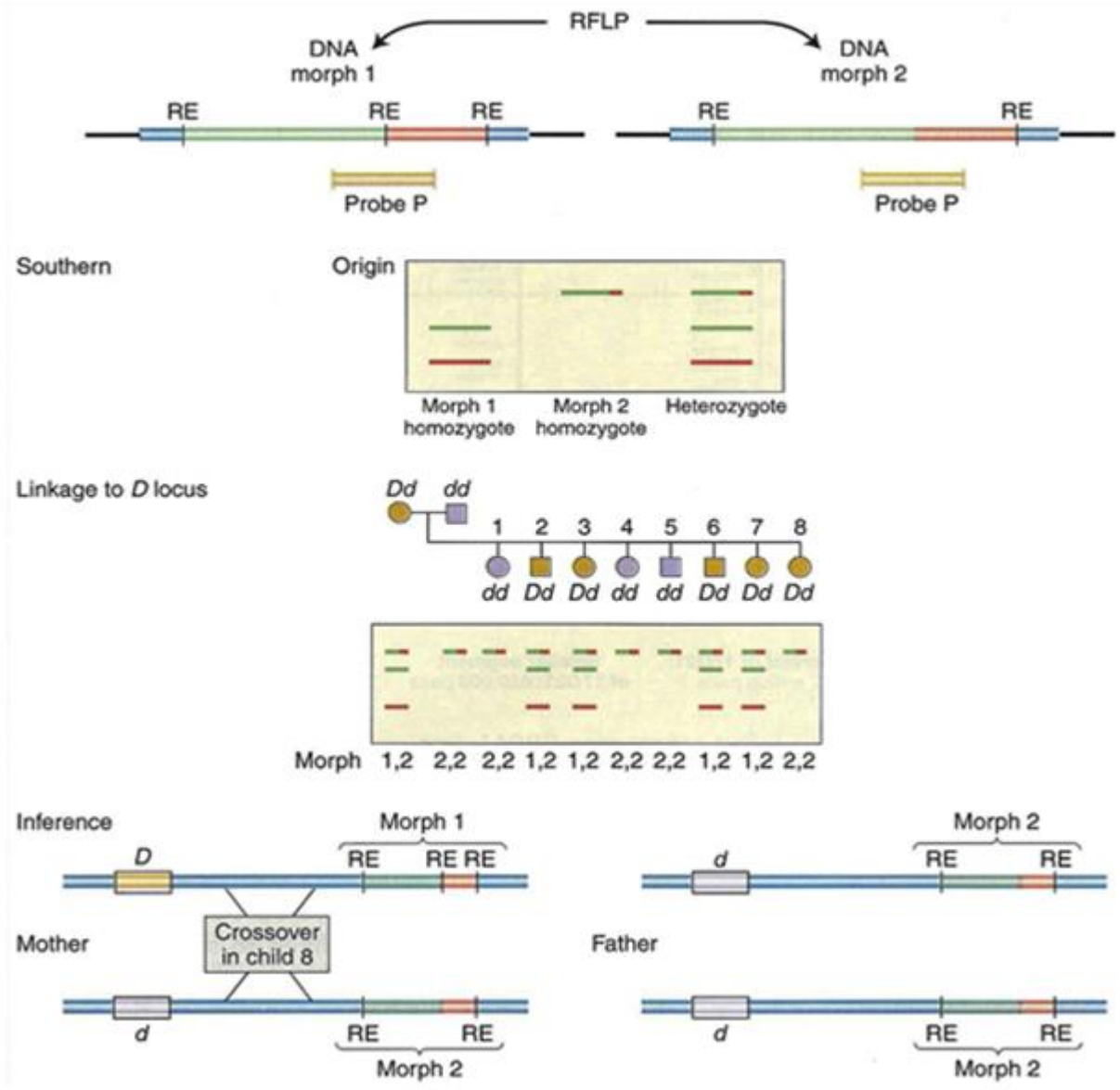
- The starting materials for this example are samples of cloned and purified β-globin alleles.
- To determine if a person is a heterozygous carrier of the mutant β-globin allele, we would directly compare the genomic DNA from that person with DNA from both a person who has sickle cell anaemia (homozygous recessive) and a person who is homozygous dominant). (Refer to Page 14 and 15)
  - In this case, since many restriction fragments will be produced upon digestion, Southern blotting, which combines gel electrophoresis and nucleic acid hybridisation, will be carried out. This allows for detection of bands which include parts of β-globin allele

(ii) **Other diseases** (RFLP closely linked to allele)

- When the gene responsible for causing the disease has not yet been identified, the presence of a mutant allele can be determined if a closely linked RFLP has been found.
- Therefore the RFLP marker and gene will almost always be inherited together.
- Alleles for Cystic Fibrosis and Huntington's Disease were first detected in this manner.




**RFLP as markers for disease-causing alleles**




**Detection and inheritance of RFLP**

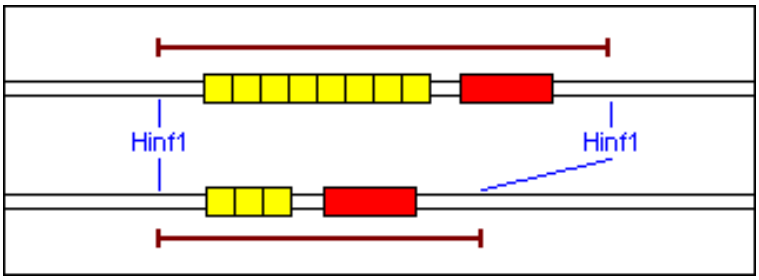
*(The pedigree of the dominant disease of the dominant disease phenotype D shows linkage of the D gene locus to the RFLP.)*

(b) **DNA Fingerprinting**

DNA fingerprinting is a technique used in **forensic science** and **paternity testing**. It utilises RFLP analysis to identify the guilty individual with a high degree of certainty because the restriction band pattern of every person is unique (except for identical twins).

Most of the polymorphisms is due to **variable number of tandem repeats (VNTR) present in the intergenic regions of the genome**.

- The **number and position of the bands formed on the gel** can be referred to as the **"DNA fingerprint"** of the individual. This allows researchers to determine if DNA samples from the same person, related people or non-related people.




**RFLP resulting from VNTR in the intergenic regions**

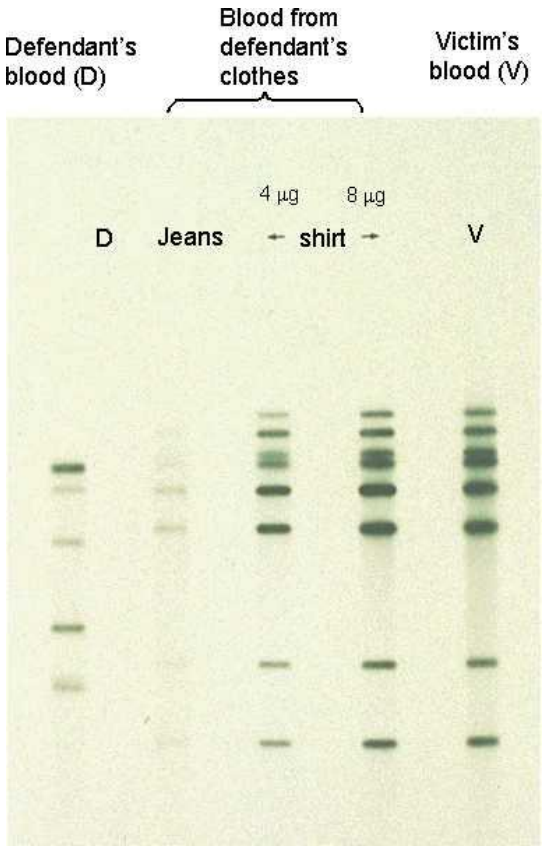
- In very rare instances, two individuals might have the same pattern of fragments when one restriction enzyme is used.
  - This problem is overcome by using a **combination of restriction enzymes**.
  - When several enzymes are used, the probability of two individuals having the same band pattern is very low, as rare as one in a billion chance.

(i) **Forensic Science**

In violent crimes, if a considerable amount of blood or tissue is available, these materials can be used to determine the blood or tissue type. However, this only allows for elimination of suspects but do not allow identification of the perpetrator. DNA fingerprinting, however, requires only a tiny amount of DNA samples to detect differences in DNA sequences (RFLPs).

To circumvent the issue of having similar RFLP profiles among individuals if only a single restriction enzyme is added, many restriction enzymes and probes are usually used in criminal investigations.

- In this example, some DNA samples were collected from a murder case. DNA samples of the victim and defendant were also collected.
  - DNA samples were subjected to restriction enzyme digestion followed by Southern blotting using radioactive probes.
  - DNA bands resulting from gel electrophoresis were exposed to probes for several RFLP markers in succession, with the previous probe washed away before the next one is applied.
  - RFLP analysis shows that DNA in blood from defendant's clothes matches the DNA fingerprint of the victim but not that of the defendant.



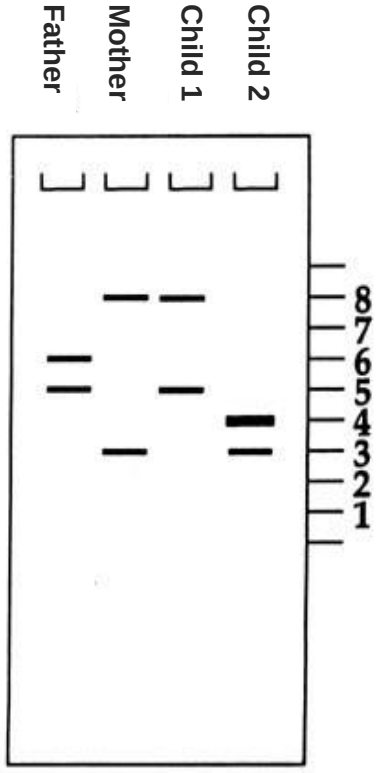

**DNA Fingerprint from a Murder Case**

(ii) **Paternity Testing**

DNA fingerprinting is used to confirm and resolve paternity issues. DNA fingerprinting can be used to identify parents of a child by **comparing the DNA fingerprints of the parents and the child**.

- A child **inherits one homologue of each pair of homologous chromosome** from his parents. Thus, his band patterns should be a **combination** of his parents' band patterns.
- When the paternity of a child is questioned, the DNA of possible children and possible parents are obtained. The children's band patterns are then matched with the adults' band patterns.
- From the band patterns in the diagram below, it can be concluded that:

  Child 1 is the child of the parents as child 1 has inherited one DNA band from each parent. Child 1 and 2 may share the same mother but have different fathers.




**DNA Fingerprint for Paternity Testing**

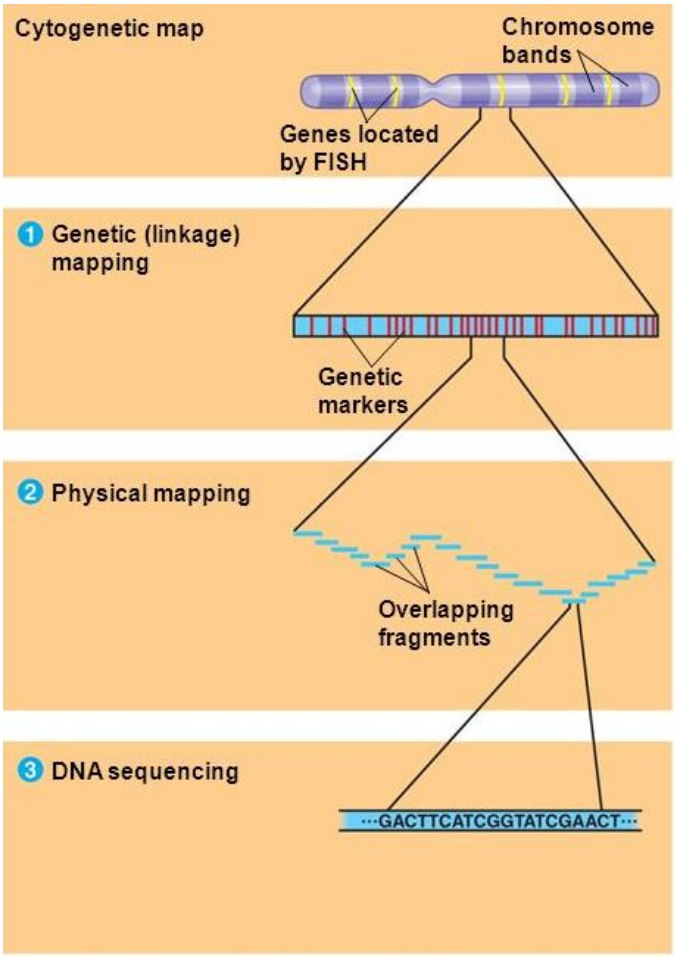
(iii) **Other possible applications**

Similarly, genetic fingerprinting is used to identify the deceased after a mass disaster by comparing DNA profiles obtained from close relatives.

(c) **Genome Mapping**

**Genome mapping** is the **mapping of genes to specific locations on chromosomes within a genome**.

There are two types of genetic maps: genetic maps and physical maps. Both genetic and physical maps **provide information of how genes and genetic markers are ordered on a chromosome**. Physical maps, however **provide the distance of how far apart the genes or markers are in terms of number of base pairs**.





**Three-Step Approach to Mapping Entire Genome**

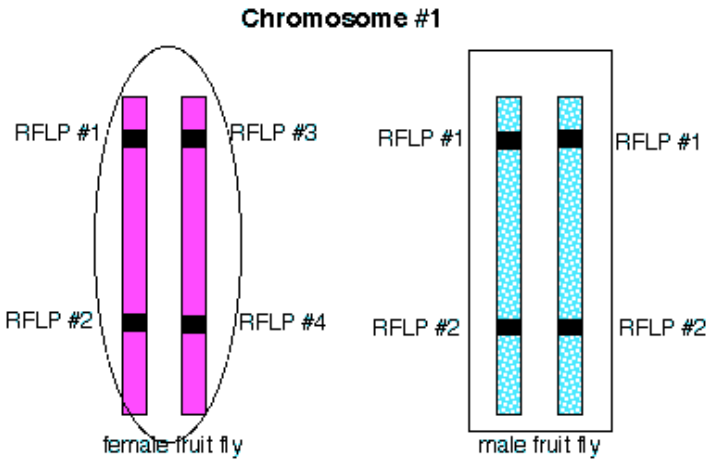
(i) **Genetic Mapping**

With cytogenetic maps of the chromosomes at hand, the initial stage in mapping the genome is to construct a genetic (linkage) map of several thousand genetic markers spaced throughout the chromosomes.

Traditionally, linkage analyses to obtain genetic maps were performed by observing phenotypes of parental and offspring generation.

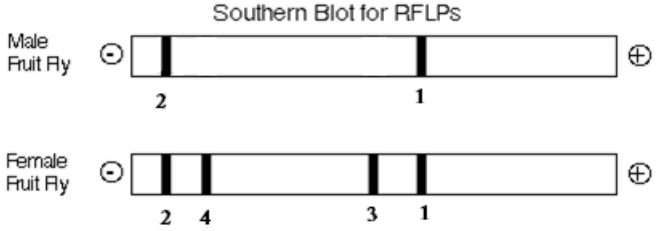
It is now possible to conduct RFLP analysis to **determine whether two RFLP loci are linked or not, and if they are, how closely they are linked**.

- A **genetic marker** is a **segment of DNA that is found at a specific site along a chromosome** and it has properties that enable it to be uniquely recognized.
- As RFLP markers are **inherited in a Mendelian fashion**, they can serve as genetic markers for a particular location (locus) in the genome
- The **frequency with which genetic markers are inherited together** is a measure of the **closeness of the two loci on a chromosome**.
- The following is used as an example to illustrate how RFLP analysis can be used to produce genetic maps. A female fly is heterozygous at two loci is mated with a male is homozygous at both loci. There are two RFLP loci with two RFLP bands possible at each locus.




**RFLP loci in Male and Female Fruit Fly Chromosome**

- When RFLP analysis is conducted, the following diagram shows the banding pattern of the parental flies in a Southern blot.

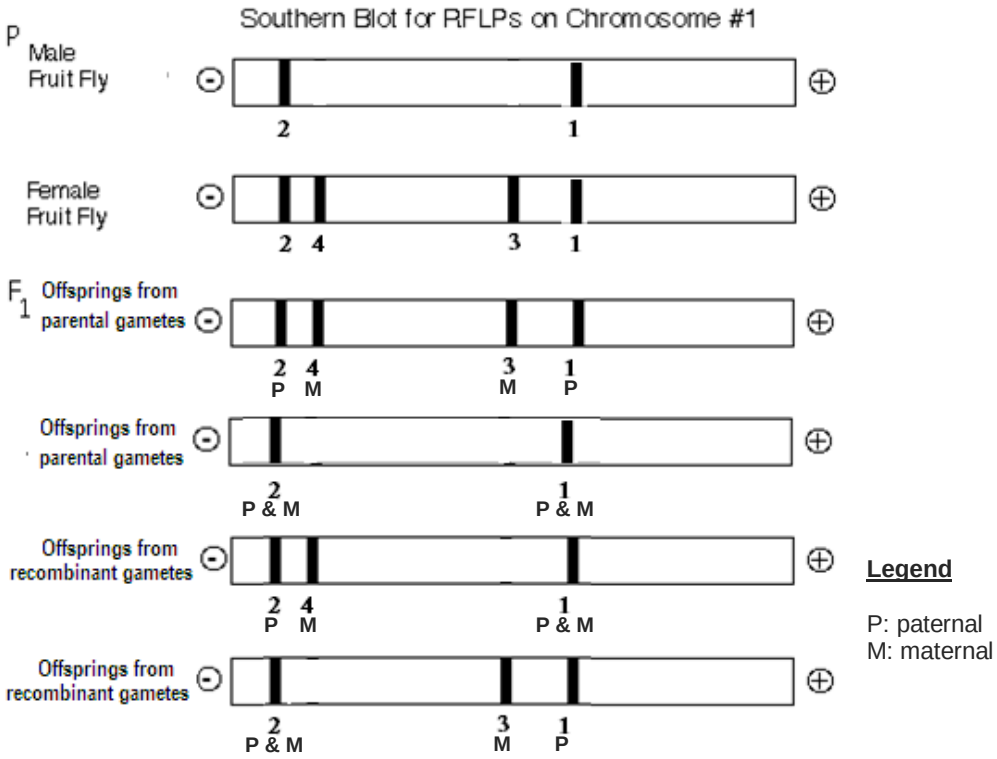



**Banding Pattern of Parental Flies in Southern Blot**

- The male can only produce one type of gamete possessing both band 1 and 2.
- The female fly can produce four different types of gametes due to the heterozygosity at the RFLP loci.
  - Parental gametes carry the exact same RFLP bands as the female parent.
  - Recombinant gametes are produced when crossing over during meiosis separates the linked RFLPs.
  - Gametes with RFLP 1 & RFLP 4 and gametes with RFLP 2 & RFLP 3 are recombinant gametes.

| Type of Gametes from Female Parent | Alleles |
|---|---|
| Parental | RFLP 1 & RFLP 2 |
| Parental | RFLP 3 & RFLP 4 |
| Recombinant | RFLP 1 & RFLP 4 |
| Recombinant | RFLP 2 & RFLP 3 |

- When these two flies mate, the frequency of the four possible progeny can be measured. From this information, the genetic distance between the upper and lower RFLP loci can be determined.



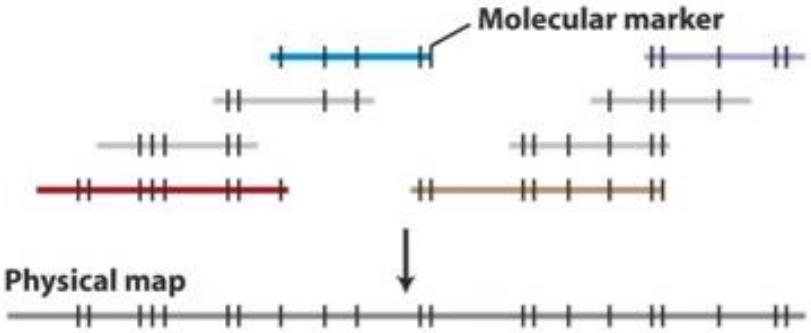

- The **closer the two RFLP loci** are, the **lower the possibility of crossing over and recombination** occurring to separate the linked RFLPs, the **lower the recombination frequency**. [X. ref to Genetics for Variation]

$$\text{Recombination frequency (\%)} = \frac{\text{number of organisms showing recombinant phenotypes}}{\text{total number of offspring}} \times 100$$

- Centimorgan (cM) is a unit for measuring genetic linkage in genetic maps. One centimorgan represents a 1% chance that a marker at one genetic locus on a chromosome will be separated from a marker at a second locus due to crossing over in a single generation.

(ii) **Physical Mapping**

The **entire genome is cloned to produce a set of overlapping fragments** that together encompass the genome. The genetic markers are ordered to form contiguous or overlapping segments of the same chromosome by comparing for overlapping sequence.




**Construction of Physical Map**

## 6. Human Genome Project

The Human Genome Project was an immense project which involved scientists from all over the world with the main aim of **mapping and sequencing the 3 billion base pairs in the human genome (22 autosomes and 2 sex chromosomes)**. It was interested in obtaining sequence of human genes and their chromosomal location and the non-coding DNA within the human genome.

- It initially started out in 1990 as a publicly funded project. The public group used a systematic method to identify genetic markers on each chromosome and created a physical map of the human genome. With the aid of these markers, the genetic information on each chromosome was sequenced.
- In 1998, a private company started to sequence the human genome using another technique called the shot-gun sequencing. In this approach, the whole human genome was digested by non-specific restriction enzymes, cloned and sequenced. The fragments contain overlapping sequences which will be pieced together using a computer program. This method proved to be more efficient as both groups completed the final draft of the human genome in 2003.

### (a) Goals

The goals of the human genome project are as follows:

- **identify** all the approximately **20,000-25,000 genes** in human DNA by 2003
- determine the **sequences** of the 3 billion chemical base pairs that make up human DNA
- store this information in databases and make it freely and totally accessible
- improve tools for data analysis, especially for sequencing and comparison of genomes of other organisms
- transfer related technologies to the private sector and improve training and manpower
- address the ethical, legal, and social issues that may arise from the project

### (b) Benefits

The completion of the Human Genome Project (HGP) brought about the following benefits:

- Molecular Medicine
  - Improved diagnosis of disease

    A detailed genetic and physical map has made it easier for researchers to locate genes in order to study their functions. Genes associated with genetic conditions and the types of proteins encoded by these genes and their interaction in complex genetic diseases can be investigated.
  - Earlier detection of genetic predispositions to disease

    HGP data can be used to compare gene sequences between diseased and healthy individuals. Hence mutations that contribute to disease can be identified and used as markers. The cloning and sequencing of disease-causing alleles is expected to play important roles in the future diagnosis and treatment of disease.
  - Rational drug design
  - Gene therapy and control systems for drugs
  - Pharmacogenomics "custom drugs"

- DNA Forensics (Identification)
  - Identify potential suspects whose DNA may match evidence left at crime scenes
  - Exonerate persons wrongly accused of crimes
  - Identify crime and catastrophe victims
  - Establish paternity and other family relationships
  - Identify endangered and protected species as an aid to wildlife officials (could be used for prosecuting poachers)
  - Detect bacteria and other organisms that may pollute air, water, soil, and food
  - Match organ donors with recipients in transplant programs
  - Determine pedigree for seed or livestock breeds
  - Authenticate consumables such as caviar and wine

- Risk Assessment
  - Assess health damage and risks caused by radiation exposure, including low-dose exposures
  - Assess health damage and risks caused by exposure to mutagenic chemicals and cancer-causing toxins
  - Reduce the likelihood of heritable mutations

- Bioarchaeology, Anthropology, Evolution, and Human Migration
  - Study evolution through germline mutations in lineages
  - Study migration of different population groups based on female genetic inheritance
  - Study mutations on the Y chromosome to trace lineage and migration of males
  - Compare breakpoints in the evolution of mutations with ages of populations and historical events

    HGP data can be used in comparative studies with other model organisms so that disease development can be investigated using other model organisms. Studying organisms *like Escherichia coli*, yeast, fruitfly and mouse helps researchers gain more insights about homologous genes, encoding proteins with similar functions in the human genome.

- Energy and Environmental Applications
  - Use microbial genomics research to create new energy sources (biofuels)
  - Use microbial genomics research to develop environmental monitoring techniques to detect pollutants
  - Use microbial genomics research for safe, efficient environmental remediation
  - Use microbial genomics research for carbon sequestration

- Agriculture, Livestock Breeding, and Bioprocessing
  - Disease-, insect-, and drought-resistant crops
  - Healthier, more productive, disease-resistant farm animals
  - More nutritious produce
  - Biopesticides
  - Edible vaccines incorporated into food products
  - New environmental cleanup uses for plants like tobacco

*Refer to the following website for elaboration on benefits of HGP:*
*http://www.ornl.gov/sci/techresources/Human_Genome/project/benefits.shtml*

(c) **Ethical Issues**

One major concern of human genome project is the ethical implication. Ethical issues arising from the use of human genome project information falls into a wide spectrum.

- Fairness in the use of genetic information by insurers, employers, courts, schools, adoption agencies, and the military, among others
  - With the mapping of genome and discovery of genes linked to many diseases and behaviours, the **fear of genetic discrimination by society is prevalent**.
  - There are concerns that **individuals possessing genes that are more prone to cancer, heart diseases or any other aliments will be made to pay a higher premium for insurance**.
  - There is much fear that **employers may discriminate against workers based on their genetic fingerprint**. They may **not employ people who are more prone to diseases**.
  - There are now legislations in United States to protect employees against this type of discrimination.
- Privacy and confidentiality of genetic information
- Psychological impact and stigmatization due to an individual's genetic differences
- Reproductive issues including adequate informed consent for complex and potentially controversial procedures / use of genetic information in reproductive decision making / reproductive rights
  - In **pre-natal testing**, DNA sample from the foetus can be extracted and screened for presence of genetic disorders. Parents might then have to **make a difficult decision of keeping or aborting the abnormal foetus**. There is a question **if this actually amounts to 'murder' since the foetus is a living organism**.
  - **Pre-implantation genetic testing** (zygotes from *in vitro* fertilization) can now be **carried out to 'check' for any genetic disorders in zygote** of parents who are at risk of passing on genetic diseases. This is viewed as **tampering with nature among many religious groups**.

- Clinical issues including the education of doctors and other health service providers, patients, and the general public in genetic capabilities, scientific limitations, and social risks / implementation of standards and quality-control measures in testing procedures
- Uncertainties associated with gene tests for susceptibilities and complex conditions (e.g., heart disease) linked to multiple genes and gene-environment interactions
- Conceptual and philosophical implications regarding human responsibility, free-will versus genetic determinism / concepts of health and disease
- Health and environmental issues concerning genetically modified foods (GM) and microbes
- Commercialization of products including property rights (patents, copyrights, and trade secrets) and accessibility of data and materials

*Refer to the following website for elaboration on ethical concerns arising from HGP:*
*http://www.ornl.gov/sci/techresources/Human_Genome/elsi/elsi.shtml*