Human Genome Project and DNA Fingerprinting

Objectives

After going through this lesson, the learners will be able to understand the following:

- Human Genome Project
- DNA Fingerprinting

Content Outline

- Introduction
- Human Genome Project (HGP)
- Goals of HGP
- Methodologies of HGP
- Salient Features of Human Genome
- Applications and Future Challenges
- DNA Fingerprinting
- Techniques for DNA Fingerprinting
- Applications of DNA Fingerprinting
- Summary

Introduction

In the preceding sections you have learnt that it is the sequence of bases in DNA that determines the genetic information of a given organism. In other words, genetic make-up of an organism or an individual lies in the DNA sequences. If two individuals differ, then their DNA sequences should also be different, at least at some places. These assumptions led to the quest of finding out the complete DNA sequence of the human genome. With the establishment of genetic engineering techniques where it was possible to isolate and clone any piece of DNA and availability of simple and fast techniques for determining DNA sequences, a very ambitious project of sequencing human genome was launched in the year 1990.

"*The Human Genome Project (HGP) was one of the great feats of exploration in history - an inward voyage of discovery rather than an outward exploration of the planet or the cosmos; an international research effort to sequence and map all of the genes - together known as the genome - of members of our species,* Homo sapiens. *Completed in April 2003, the HGP gave*

us the ability, for the first time, to read nature's complete genetic blueprint for building a human being". …….. National Human Genome Research Institute

Human Genome Project

The Human Genome Project (HGP) was called a mega project. You can imagine the magnitude and the requirements for the project if we simply define the aims of the project as follows:

- Human genome is said to have approximately 3 x 10^9 bp, and if the cost of sequencing required is US $ 3 per bp (the estimated cost in the beginning), the total estimated cost of the project would be approximately 9 billion US dollars.
- Further, if the obtained sequences were to be stored in typed form in books, and if each page of the book contained 1000 letters and each book contained 1000 pages, then 3300 such books would be required to store the information of DNA sequence from a single human cell.
- The enormous amount of data expected to be generated also necessitated the use of high speed computational devices for data storage and retrieval, and analysis.
- HGP was closely associated with the rapid development of a new area in biology called Bioinformatics.

Goals of HGP

Some of the important goals of HGP were as follows:

- Identify all the approximately 20,000-25,000 genes in human DNA;
- Determine the sequences of the 3 billion chemical base pairs that make up human DNA;
- Store this information in databases;
- Improve tools for data analysis;
- Transfer related technologies to other sectors, such as industries;
- Address the ethical, legal, and social issues (ELSI) that may arise from the project.

The Human Genome Project was a 13-year project coordinated by the U.S. Department of Energy and the National Institute of Health. During the early years of the HGP, the Wellcome Trust (U.K.) became a major partner; additional contributions came from Japan, France, Germany, China and others. The project was completed in 2003. Knowledge about the effects of DNA variations among individuals can lead to revolutionary new ways to diagnose, treat and someday prevent the thousands of disorders that affect human beings. Besides providing

clues to understanding human biology, learning about non-human organisms DNA sequences can lead to an understanding of their natural capabilities that can be applied toward solving challenges in health care, agriculture, energy production, environmental remediation. Many non-human model organisms, such as bacteria, yeast, *Caenorhabditis elegans* (a free living non-pathogenic nematode), *Drosophila* (the fruit fly), plants (rice and *Arabidopsis*), etc., have also been sequenced.

Methodologies

The methods involved two major approaches. One approach focused on identifying all the genes that are expressed as RNA (referred to as Expressed Sequence Tags (ESTs). The other took the blind approach of simply sequencing the whole set of genome that contained all the coding and non-coding sequence, and later assigning different regions in the sequence with functions (a term referred to as Sequence Annotation).

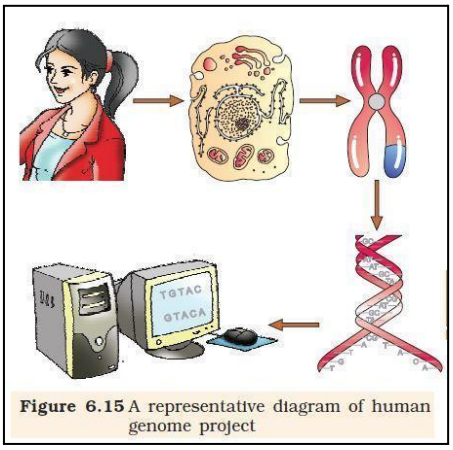


Figure 6.15 A representative diagram of human genome project

For sequencing, the total DNA from a cell is isolated and converted into random fragments of relatively smaller sizes (recall DNA is a very long polymer, and there are technical limitations in sequencing very long pieces of DNA) and cloned in a suitable host using specialised vectors. The cloning resulted in amplification of each piece of DNA fragment so that it subsequently could be sequenced with ease.

The commonly used cloning hosts were bacteria and yeast, and the vectors were called BAC (bacterial artificial chromosomes), and YAC (yeast artificial chromosomes). The fragments were sequenced using automated DNA sequencers that worked on the principle of a method developed by Frederick Sanger. (Remember, Sanger is also credited for developing method for determination of amino acid sequences in proteins). These sequences were then arranged based on some overlapping regions present in them. This required generation of overlapping

fragments for sequencing. Alignment of these sequences was humanly not possible. Therefore, specialized computer based programs were developed. These sequences were subsequently annotated and were assigned to each chromosome. The sequence of chromosome 1 was completed only in May 2006 (this was the last of the 24 human chromosomes – 22 autosomes and X and Y – to be sequenced). Another challenging task was assigning the genetic and physical maps on the genome. This was generated using information on polymorphism of restriction endonuclease recognition sites, and some repetitive DNA sequences known as microsatellites (one of the applications of polymorphism in repetitive DNA sequences shall be explained in next section of DNA fingerprinting).

Salient Features of Human Genome

Some of the salient observations drawn from human genome project are as follows:

- The human genome contains 3164.7 million nucleotide bases.
- The average gene consists of 3000 bases, but sizes vary greatly, with the largest known human gene being dystrophin at 2.4 million bases.
- The total number of genes is estimated at 30,000–much lower than previous estimates of 80,000 to 1,40,000 genes. Almost all (99.9 per cent) nucleotide bases are exactly the same in all people.
- The functions are unknown for over 50 per cent of the discovered genes.
- Less than 2 percent of the genome codes for proteins.
- Repeated sequences make up a very large portion of the human genome.
- Repetitive sequences are stretches of DNA sequences that are repeated many times, sometimes hundred to thousand times. They are thought to have no direct coding functions, but they shed light on chromosome structure, dynamics and evolution.
- Chromosome 1 has the most genes (2968), and the Y has the fewest (231).
- Scientists have identified about 1.4 million locations where single base DNA differences (SNPs – single nucleotide polymorphism, pronounced as ‘snips’) occur in humans.

This information promises to revolutionise the processes of finding chromosomal locations for disease-associated sequences and tracing human history.

Applications and Future Challenges

Deriving meaningful knowledge from the DNA sequences will define research through the coming decades leading to our understanding of biological systems. This enormous task will require the expertise and creativity of tens of thousands of scientists from varied disciplines

in both the public and private sectors worldwide. One of the greatest impacts of having the HG sequence may well be enabling a radically new approach to biological research. In the past, researchers studied one or a few genes at a time. With whole-genome sequences and new high-throughput technologies, we can approach questions systematically and on a much broader scale. They can study all the genes in a genome, for example, all the transcripts in a particular tissue or organ or tumor, or how tens of thousands of genes and proteins work together in interconnected networks to orchestrate the chemistry of life.

DNA Fingerprinting

As stated in the preceding section, 99.9 percent of base sequence among humans is the same. *Assuming human genome as 3 × 10^9 bp, in how many base sequences would there be differences?*

It is these differences in sequence of DNA which make every individual unique in their phenotypic appearance. If one aims to find out genetic differences between two individuals or among individuals of a population, sequencing the DNA every time would be a daunting and expensive task. Imagine trying to compare two sets of 3 × 10^6 base pairs. DNA fingerprinting is a very quick way to compare the DNA sequences of any two individuals.

DNA fingerprinting involves identifying differences in some specific regions in a DNA sequence called repetitive DNA, because in these sequences, a small stretch of DNA is repeated many times. These repetitive DNA are separated from bulk genomic DNA as different peaks during density gradient centrifugation. The bulk DNA forms a major peak and the other small peaks are referred to as satellite DNA. Depending on base composition (A : T rich or G:C rich), length of segment, and number of repetitive units, the satellite DNA is classified into many categories, such as micro-satellites, mini-satellites etc. These sequences normally do not code for any proteins, but they form a large portion of the human genome. These sequences show a high degree of polymorphism and form the basis of DNA fingerprinting. Since DNA from every tissue (such as blood, hair-follicle, skin, bone, saliva, sperm etc.), if an individual shows the same degree of polymorphism, they become a very useful identification tool in forensic applications. Further, as the polymorphisms are inheritable from parents to children, DNA fingerprinting is the basis of paternity testing, in case of disputes.

As polymorphism in DNA sequence is the basis of genetic mapping of human genome as well as of DNA fingerprinting, it is essential that we understand what DNA polymorphism

means in simple terms. Polymorphism (variation at genetic level) arises due to mutations. (Recall different kinds of mutations and their effects that you have already studied in Chapter 5, and in the preceding sections in this chapter.) New mutations may arise in an individual either in somatic cells or in the germ cells (cells that generate gametes in sexually reproducing organisms). If a germ cell mutation does not seriously impair individual's ability to have offspring who can transmit the mutation, it can spread to the other members of population (through sexual reproduction). Allelic (again recall the definition of alleles from Chapter 5) sequence variation has traditionally been described as a DNA polymorphism if more than one variant (allele) at a locus occurs in human population with a frequency greater than 0.01. In simple terms, if an inheritable mutation is observed in a population at high frequency, it is referred to as DNA polymorphism. The probability of such variation to be observed in noncoding DNA sequences would be higher as mutations in these sequences may not have any immediate effect/impact on an individual's reproductive ability. These mutations keep on accumulating generation after generation, and form one of the basis of variability/polymorphism.

There are a variety of different types of polymorphisms ranging from single nucleotide change to very large scale changes. For evolution and speciation, such polymorphisms play a very important role, and you will study these in detail at higher classes.

The technique of DNA Fingerprinting was initially developed by Alec Jeffreys. He used a satellite DNA as a probe that showed a very high degree of polymorphism. It was called the Variable Number of Tandem Repeats (VNTR).

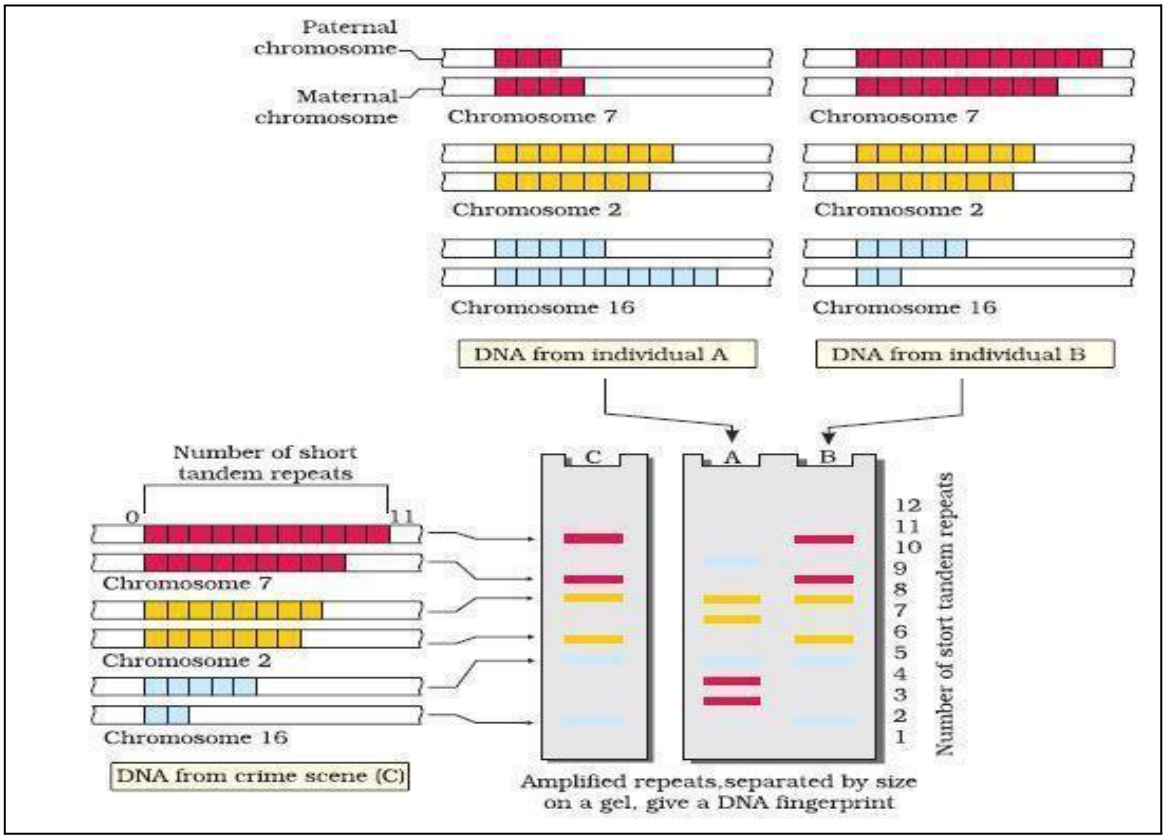
Techniques for DNA Fingerprinting

The technique, as used earlier, involved Southern blot hybridisation using radio labelled VNTR as a probe. It included

- Isolation of DNA,
- Digestion of DNA by restriction endonucleases,
- Separation of DNA fragments by electrophoresis,
- Transferring (blotting) of separated DNA fragments to synthetic membranes, such as nitrocellulose or nylon,
- Hybridisation using labelled VNTR probe, and
- Detection of hybridised DNA fragments by autoradiography.

The VNTR belongs to a class of satellite DNA referred to as mini-satellite. A small DNA sequence is arranged tandemly in many copy numbers. The copy number varies from

chromosome to chromosome in an individual. The numbers of repeats show a very high degree of polymorphism. As a result the size of VNTR varies in size from 0.1 to 20 kb.



A schematic representation of DNA fingerprinting is shown in Figure

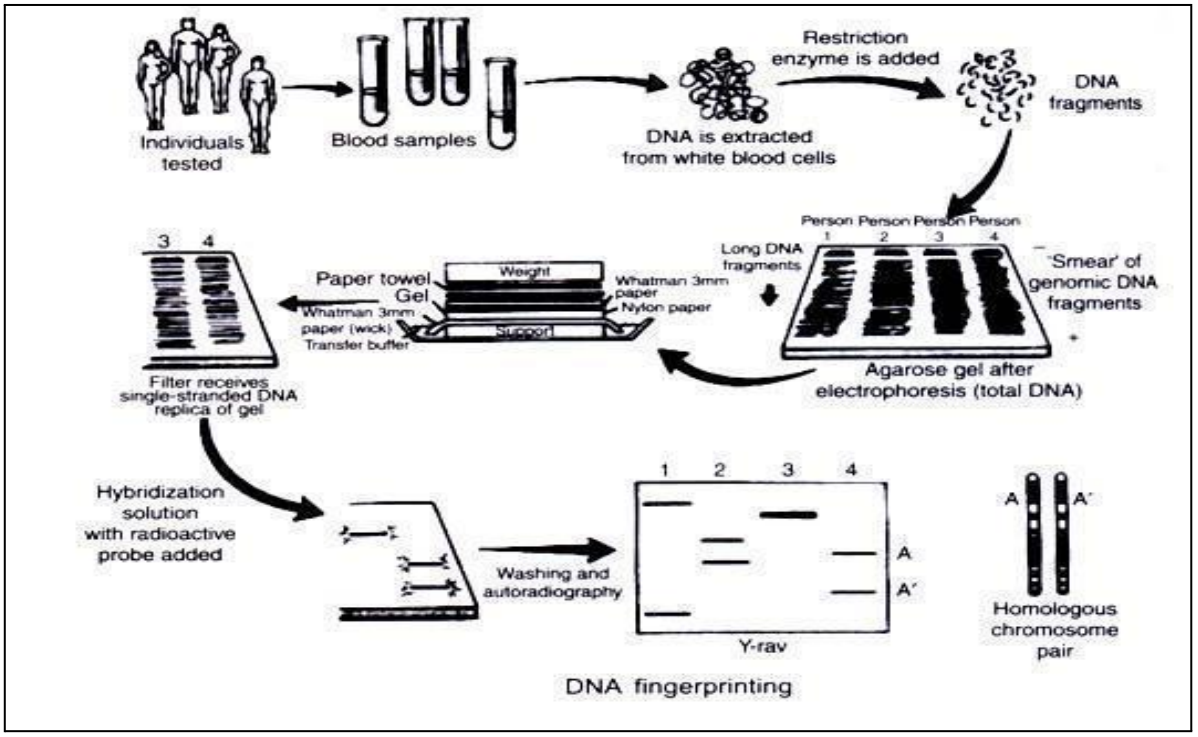
Consequently, after hybridisation with VNTR probe, the autoradiogram gives many bands of differing sizes. These bands give a characteristic pattern for an individual DNA (Figure 6.16). It differs from individual to individual in a population except in the case of monozygotic (identical) twins. The sensitivity of the technique has been increased by use of polymerase chain reaction (PCR–you will study about it in Chapter 11).Consequently, DNA from a single cell is enough to perform DNA fingerprinting analysis. In addition to application in forensic science, it has much wider application, such as in determining population and genetic diversities. Currently, many different probes are used to generate DNA fingerprints.

The other types of DNA fingerprinting methods in use at this time are RFLP, Amp FLP and STR.

RFLP

Restriction fragment length polymorphism (RFLP) analyzes the length of the strands of the DNA molecules with repeating base pair patterns. Inside the nucleus of each human cell, the chromosome contains the long strands of DNA molecules.

Each DNA strand contains a large number of coding genes while DNA is non coding. However the 95% non-coding genes contain identifiable repetitive sequences of base pairs, which are known as VNTR. As the length of repeats varies from person to person, digestion of VNTP with RES provides fragments of different lengths to different individuals. This is known as Restriction Fragment Length Polymorphism (RFLP).



The restriction fragment length polymorphism analysis is used to detect the repeated sequences by determining a specific pattern to the VNTR, which becomes the person's DNA fingerprint. Steps include, isolation of DNA, digestion of DNA by restriction endonucleases, separation of DNA fragments by electrophoresis, transferring (blotting) of separated DNA fragments to synthetic membranes.

AmpFLP or AFLP

AmpFLP (Amplified fragment length polymorphism) AmpFLP came into vogue in the 90's and is still popular in the smaller countries involved in the process of DNA fingerprinting. It is a relatively less complicated operation and has the cost-effectiveness of the procedure.

By using the PCR analysis to amplify the minisatellite loci of the human cell, this method proved quicker in recovery than the RFLP. There are issues of bunching of the VTRN, causing misidentifications in the process due to the use of gel in its analysis phase.

STR

The system most widely used for DNA fingerprinting is the Short tandem repeat (STR) methodology.

The STR analyzes how many times base pairs repeat themselves on a particular location on a strand of DNA. The DNA comparisons can match the possibilities into an almost endless range; therefore, it is the big advantage in this method.

DNA fingerprinting has been extremely successful for use in the personal identification of criminal suspects. DNA testing for ethnicity, identification of the deceased, as well as court-approved paternity tests. However, Still DNA poses issues as the VNTRs are not evenly distributed in all people as they are inherited. Further, there is still the imperfect human element as the final voice in the administration of all DNA fingerprinting procedures.

Applications of DNA Fingerprinting

Paternity and Maternity

Because a person inherits his or her VNTRs from his or her parents, VNTR patterns can be used to establish paternity and maternity. The patterns are so specific that a parental VNTR pattern can be reconstructed even if only the children's VNTR patterns are known (the more children produced, the more reliable the reconstruction). Parent-child VNTR pattern analysis has been used to solve standard father-identification cases as well as more complicated cases of confirming legal nationality and, in instances of adoption, biological parenthood.

Criminal Identification and Forensics

DNA isolated from blood, hair, skin cells, or other genetic evidence left at the scene of a crime can be compared, through VNTR patterns, with the DNA of a criminal suspect to determine guilt or innocence. VNTR patterns are also useful in establishing the identity of a homicide victim, either from DNA found as evidence or from the body itself.

Personal Identification

The notion of using DNA fingerprints as a sort of genetic barcode to identify individuals has been discussed, but this is not likely to happen anytime in the foreseeable future. The

technology required to isolate, keep on file, and then analyze millions of much specified VNTR patterns is both expensive and impractical. Social security numbers, picture ID, and other more mundane methods are much more likely to remain the prevalent ways to establish personal identification. DNA fingerprinting can be used to study diversity at genetic level. Useful in taxonomy and evolutionary studies.

Summary

Human genome project was a mega project that aimed to sequence every base in the human genome. This project has yielded much new information. Many new areas and avenues have opened up as a consequence of the project. DNA Fingerprinting is a technique to find out variations in individuals of a population at DNA level. It works on the principle of polymorphism in DNA sequences. It has immense applications in the field of forensic science, genetic biodiversity and evolutionary biology.