MANE-001
Human Genetics

Indira Gandhi
National Open University
School of Social Sciences

Block

5

HUMAN MOLECULAR GENETICS

Expert Committee

Prof. P. Dash Sharma (Retd.)
Dept. of Anthropology
Ranchi University, Ranchi

Prof. P. Veerraju (Retd.)
Dept. of Human Genetics
Andhra University
Visakhapatnam

Prof. S.M.S. Chahal
Dept. of Human Biology
Punjabi University, Patiala

Prof. A. Papa Rao
Dept. of Anthropology
Sri Venkateswara University
Tirupati

Dr. Roli Mathur
Scientist 'C'
Division of Basic Medical Sciences
Indian Council of Medical Research
New Delhi

Dr. Seema Kalra
Assistant Professor
Dept. of Biochemistry
IGNOU, New Delhi

Faculty of Anthropology
SOSS, IGNOU

Dr. Rashmi Sinha, Reader
Discipline of Anthropology
IGNOU, New Delhi

Dr. P. Venkatramana
Assistant Professor
Discipline of Anthropology
IGNOU, New Delhi

Dr. Rukshana Zaman
Assistant Professor
Discipline of Anthropology
IGNOU, New Delhi

Dr. Mitoo Das
Assistant Professor
Discipline of Anthropology
IGNOU, New Delhi

Dr. K. Anil Kumar
Assistant Professor
Discipline of Anthropology
IGNOU, New Delhi

Academic Assistance provided by
Dr. N.K. Mungreiphy, Research Associate
(DBT) for the Expert Committee meeting

Programme Coordinator: Dr. Rashmi Sinha, SOSS, IGNOU, New Delhi

Course Coordinator: Dr. P. Venkatramana, SOSS, IGNOU, New Delhi

Content Editor

Prof. P.B. Gai
Department of Applied Genetics
Karnataka University, Dharwad, Karnataka State

Block Preparation Team

Unit Writers

Dr. V. Lakshmi (Unit 1)
Assistant Professor, Department of Human Genetics
Andhra University, Visakhapatnam

Dr. A. Chandrasekhar (Unit 2)
Anthropological Survey of India
Southern Regional Centre, Mysore

Prof. T. Padma (Retd.) (Unit 3)
Department of Genetics
Osmania University, Hyderabad

Cover Design

Dr. Mitoo Das
Asstt. Professor
Anthropology
SOSS, IGNOU
New Delhi

Authors are responsible for the academic content of this course as far as the copyright issues are concerned.

Print Production

Mr. Manjit Singh
Section Officer (Pub.), SOSS, IGNOU, New Delhi

July, 2012

ISBN-978-81-266-6137-4

Further information on Indira Gandhi National Open University courses may be obtained from the University's office at Maidan Garhi. New Delhi-110 068.

Printed and published on behalf of the Indira Gandhi National Open University, New Delhi by the Director, School of Social Sciences.

Laser Typeset by : Tessa Media & Computers, C-206, A.F.E.-II, Okhla, New Delhi

Printed at :

BLOCK 5 HUMAN MOLECULAR GENETICS

Molecular genetics studies the structure and function of genes at molecular level. Research in Molecular genetics employs both the methods of genetics and molecular biology. Application of molecular genetics includes use of molecular information to determine the patterns of descent and understanding genetic mutations that can cause certain types of diseases. Using the techniques of molecular genetics we can discover the reasons why traits are carried on and how and why some may mutate.

The core of Molecular genetics is the structure and function, variations (Polymorphisms), causes for variations in DNA sequences; inter relation between the DNA and RNA molecules, curative measures in default structure and function of genetic material. Many discoveries were made by Human genome project in the areas of human molecular genetics. The draft of Human genome sequence was published in 2001 and the finished version published in 2003. It has thrown light on cell and development, molecular phylogentics, Mendelian genetics, genetics of complex diseases and pharmaco genetics.

Application of Human molecular genetics cover three major areas of research - Molecular genetics information in anthropological and historical research useful in tracing human origins, prehistoric migrations and demographic history; the accuracy of DNA based tests to identify people in Forensic analysis; and Molecular genetics information used to identify genetic variants of multi-factorial and complex diseases, which can explain differences between human populations with respect to increased susceptibility or resistance to certain diseases.

This block consists of 3 units. Unit 1 introduces you to Molecular Genetics and the second unit describes on DNA polymorphisms and their use in variation, evolution and disease manifestation. The third unit deals with Human Genome Project.

ignou
THE PEOPLE'S
UNIVERSITY

UNIT 1 INTRODUCTION TO MOLECULAR GENETICS

Contents

Learning Objectives

After reading this unit, you would be able to:

- discuss what is molecular genetics and how is it useful to mankind;
- describe structure and function of nucleic acids;
- imagine how the genome is organised in humans; and
- explain how genes are expressed and the ways in which gene expression is regulated.

1.1 INTRODUCTION

DNA is the master molecule which carries the genetic information from one generation to the other. Study of Molecular Genetics accelerated since April 25th, 1953 when James Watson and Francis Crick proposed the structure of DNA which was published in Journal called 'Nature'. Molecular Genetics deals with the flow of genetic information and its regulation. In simple terms it can be defined as the field of biology which studies the structure and function of genes at molecular level.

1.2 SCOPE OF MOLECULAR GENETICS

Development of techniques like nucleic acid hybridisation, cloning, sequencing etc. brought a revolutionary change in Molecular Genetics. It is of major interest to the students of biology and medicine. Though it has a lot of significance in many fields, we'll confine ourselves to the applications of molecular genetics to the mankind. They are as follows:

i) Diagnosis of infectious diseases**:** Normally microorganisms are detected in the laboratory using biochemical methods. In case of molecular techniques, microorganisms are detected by using probes (short DNA or RNA sequence) which are complementary to a part of genome of the microbe. The advantage of using molecular methods is:

- Identification of pathogen is done within a short time;
- No need to cultivate the microbes;
- Latent infections can also be identified when no antibody is formed; and
- The technique can be used even when the microorganism cannot be cultured.

ii) Diagnosis of genetic diseases : Before the advent of the above techniques, counselors used to give risk estimate like one fourth risk of getting the disease, if the parents are heterozygous for an autosomal trait. But now by directly testing for the mutation, they are able to confirm the presence or absence of mutation in the fetus. It is of immense help in prenatal diagnosis.

iii) Individuals can be identified and relationship can be determined by DNA fingerprinting.

iv) Mouse models for genetic diseases have been developed by creating transgenic mice.

v) Production of vaccines, antibodies and therapeutic proteins using recombinant DNA technology. Eg; Insulin, Human Growth Hormone.

vi) It has great potential for treating disease. Gene therapy is a process where the cells of a patient are genetically modified to alleviate disease.

vii) With the development of recombinant DNA technology, identification of disease genes became much easier. Once the disease gene is identified, a molecular test can be designed for diagnosis of genetic disease.

DNA Fingerprinting

Just like no two individuals have identical finger prints, no two individuals have identical genetic information, except monozygotic twins. Unlike finger prints which are present only on the tips of the fingers, the genetic information which is unique to the individual is present in each and every cell. Alec Jeffrey discovered repetitive sequences called minisatellites to be unique to every individual. DNA fingerprinting is a technique which makes use of these sequences to evaluate genetic information. It's a quick way to compare the DNA sequences of any two living organisms. It is used for personal identification, identification of the parents, when babies are switched in hospital, identification of criminals etc.

1.3 DEOXYRIBONUCLEIC ACID (DNA)

DNA is the thread of life. It is the hereditary material in all organisms except in certain RNA viruses. All the information that is needed for the development, behavior, well being etc. of an individual is encoded in its structure. The genetic information that's stored in DNA flows through RNA to proteins. This flow of genetic information is referred to as central dogma of molecular biology. Though the information is present in the DNA, it is the activity of proteins that is responsible for the inherited traits. The function of DNA is to direct its own replication and to direct transcription.

The number of DNA molecules present in a cell is equal to the number of chromosomes per cell. When compared to the length of the chromosome, the length of the DNA is very very long. It's a matter of interest, how this long DNA fits into the cell whose diameter is in microns (like a long snake fitting into a small basket). This is possible because of the winding of the DNA around histone proteins into structures called nucleosomes. Nucleosomes are further coiled and coiled to form chromosome. So each chromosome is nothing but a single DNA molecule along with proteins.

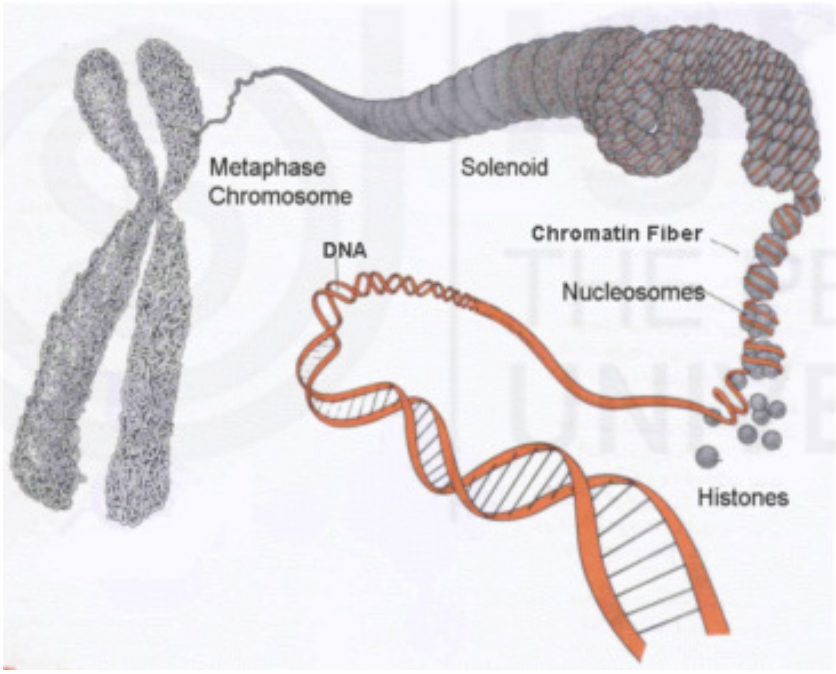


Fig.1.1: Packing of DNA into chromosomes (Source:http://www.prism.gatech.edu/~gh19/b1510/dnarep.htm)

1.3.1 Structure of DNA

The structure of DNA was proposed by Watson and Crick in the year 1953 for which they won Nobel Prize in 1962. The structure of DNA molecule resembles a gently twisted ladder. Two long polynucleotide chains represent the rails of the ladder. They're coiled around a central axis to form right handed double helix. The rungs are made up of nitrogen bases which are held together by Hydrogen bonds. The basic unit of DNA is nucleotide. It's composed of three subunits-nitrogen base, sugar and phosphate.

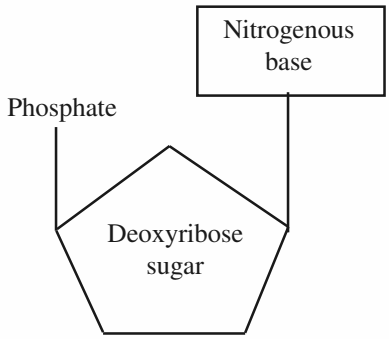


Fig. 2.2: Nucleotide

There are two kinds of bases- purines (double ringed) and pyrimidines (single ringed). A purine (Adenine and Guanine) always pairs with a pyrimidine (Thymine and Cytosine). Therefore, the amount of purines present in a DNA molecule is equal to the amount of pyrimidines. Adenine always pairs with Thymine with two Hydrogen bonds whereas Guanine always pairs with Cytosine with three Hydrogen bonds. The two strands of the DNA can be separated easily during replication because of the weak Hydrogen bonds.

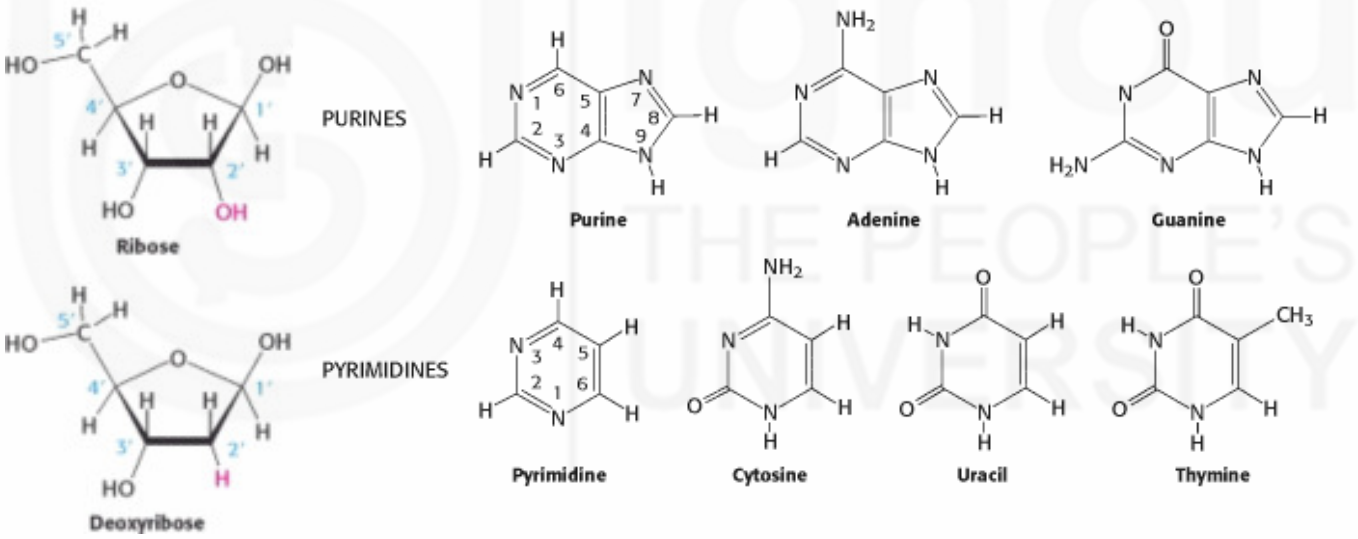


Fig. 1.3: Pentose sugar ; Purines and pyrimidines (Source : Berg JM et al, 2002 *Biochemistry* WH Freeman & Co.)

The sugar is a pentose sugar which is deoxyribose in DNA because it lacks Oxygen at second Carbon position. To distinguish between the Carbon atoms present in the base and sugar, the Carbon atoms in the sugar are given a prime (‘). The bases attach to the sugar (at 1’C) by glycosidic bond.

The phosphate group links the 3’C atom of one sugar with the 5’C atom of adjacent sugar by a phospodiester bond. Because of the negative charges present on the phosphate groups, DNA is a polyanion. *In vivo*, these charges are neutralised by the positively charged histone proteins.

1.3.2 Features of Double Helix

The two strands of the double helix are antiparallel ie., the 5’ end of one strand aligns with 3’ end of other strand.

Because of the specific base pairing, the two strands are complementary to each other which means that if we know the sequence of one strand we can infer the sequence of the other.

The bases are stacked on one another 3.4A^o apart and are perpendicular to the axis of the double helix. The diameter of the helix is about 20A^o and each complete turn of helix measures 34A^o, thus accommodating 10 base pairs in each turn.

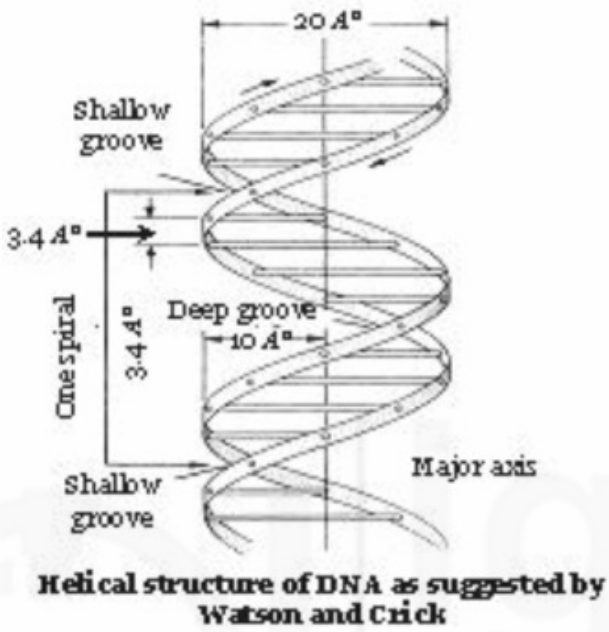


Fig. 1.4: Structure of DNA (Source:http://www.transtutors.com/chemistry-homework-help/biomolecules/dna-structure.aspx)

Stacking of base pairs (bp) results in major and minor grooves in DNA. Major groove is rich in chemical information and is recognised by sequence specific DNA binding proteins.

1.3.3 How DNA Decides the Hereditary Features

The structure of DNA is the same in all organisms with same four nitrogenous bases-A,T,G and C. Then what makes the difference between plants and animals or how does a zygote know to develop into a monkey or a human? It's the order of the base sequence that makes all the difference. It's not the same in all. The bases are present in different amounts in different species.

1.4 GENOME

A genome is the total genetic information present in a cell. Basing on the complexity of humans, if you assume that among all species, the human beings have the largest amount of DNA, you're mistaken. This is because many plant species have much more DNA per cell compared to humans. Even among vertebrates, it's the amphibians which have the greatest amount of DNA per cell. The organisation of human genome is very complex. It comprises of two genomes (Nuclear and Mitochondrial).

1.4.1 Organisation of Nuclear Genome

The nuclear genome constitutes more than 99% of the total genome. The haploid genome contains 3 billion bp. The haploid genome is distributed in 23 different types of chromosomes (22 autosomes and 1 allosome). Each chromosome contains many genes. The genes are not uniformly distributed on the chromosomes. A certain area of the chromosome may be rich in genes while areas like centromere and telomeres are largely devoid of genes. Some chromosomes are rich in genes (22^{nd} chromosome) some are gene poor (4^{th} chromosome). The genes which are part of same metabolic pathway may be on different chromosomes and genes which are no way connected to metabolic pathway may be side by side on the chromosome.

There is tremendous variation in the size of the gene, size of the exon as well as intron. On an average an exon may contain < 200bp. Size of the intron may vary from 100bp to >100,000bp. About 1.5% of the total genome is coding (Exon is the coding region and Intron is non coding).

A number of protein coding genes in the human genome form gene families. A set of genes which code for similar protein sequences or which have nucleotide sequence similarity form a gene family (just like related individuals make a family). They arose by duplication of the ancestral gene and accumulation of independent mutations over a period of time. Eg: members of beta globin gene family. An individual won't have same beta globin throughout his development. Apart from beta globin (â) there are different genes like ä, Gã, Aã and å which code for slightly different polypeptides. They express during different stages of development of an individual and forms a gene family. Members of a gene family generally appear as a cluster or they may be dispersed.

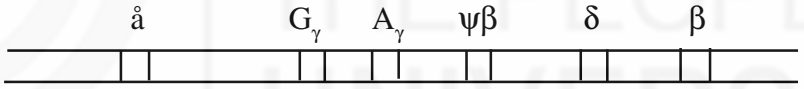


Fig. 1.5: Human beta globin gene cluster

More than half of the genome contains repetitive sequences. Basing on the number of copies per genome, the DNA sequences are classified into-unique sequences (1-10 copies), moderately repetitive sequences (10-10^5copies) and highly repetitive sequences (>10^5copies). Unique sequences include most of the genes which code for proteins. Example for moderately repetitive sequences is the genes which code for ribosomal RNA and histone proteins. Highly repetitive sequences are tandemly arranged and are transcriptionally inactive. They are once again classified into mega satellite, satellite, mini satellite and micro satellite according to the decreasing size of the repeat. Mega satellites are very few in number. Satellite DNA is present in the centromeric region of the chromosomes. The length of the mini satellite DNA is quite variable among individuals and is the basis for DNA fingerprinting. Microsatellites constitute single base runs, di, tri and tetra nucleotide repeats.

Transposons, the DNA sequences which are capable of moving from one part of the genome to other constitute about 45% of the total human genome. Most of them are nonfunctional. Transposition is RNA mediated ie., DNA is first transcribed into RNA and then reverse transcribed into cDNA which is the double stranded form is inserted elsewhere in the genome.

There are pseudogenes in the genome which are nonfunctional copies of a functional gene eg. pseudo beta(øâ) in beta globin gene family. They also arose by duplication of the ancestral gene but in course of time accumulated mutations which rendered them non functional. Few overlapping genes (in class III region of HLA complex present on 6th chromosome.) and genes within genes (presence of two genes within the intron of clotting factor VIII gene) also exist in the genome.

1.4.2 Organisation of Mitochondrial Genome

The total amount of mitochondrial genome is <1%. It varies per cell basing on the number of mitochondrial DNA(mt DNA) molecules per mitochondrion and the number of mitochondria per cell.

The size of the human mtDNA is 16,569 bp. It is circular and double stranded. It lies naked in the organelle. mtDNA isn't associated with histone proteins. It contains a light chain and a heavy chain. Heavy strand is rich in guanines and light strand is rich in cytosines. DNA is triple stranded at a region, due to duplication of a section of heavy strand and is called D loop. D loop has no coding sequences. mtDNA has altogether 37 genes out of which 13 code for polypeptides, 22 for tRNAs and 2 for rRNAs. In contrast to nuclear genome, about 93% of the mtDNA is coding. The genome is compact with no introns and presence of overlapping genes. Mitochondria have their own ribosomes on which polypeptides are synthesized. It has a slightly different genetic code when compared to that followed by the nuclear genome.

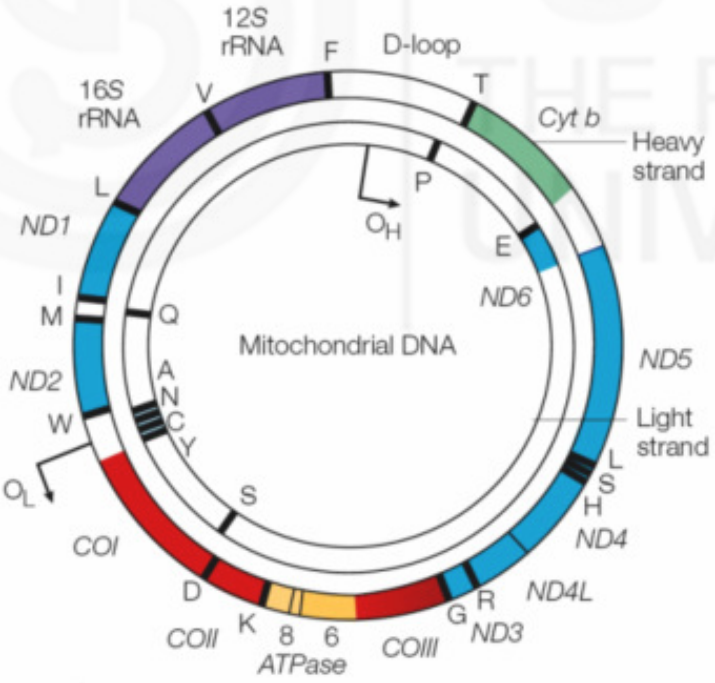


Fig.1.6: Mitochondrial genome (Source : http://www.nature.com/scitable/content/the-role-of-the-mitochondrial-genome-in-61848)

The mitochondria synthesize only some of the proteins needed by it, others being synthesized by the nuclear genes. The proteins produced by the nuclear genome are imported into mitochondria. mtDNA shows maternal inheritance because all the mitochondria received by the zygote are from the ovum. Mutations in mtDNA are responsible for certain diseases in humans.

1.5 GENETIC CODE

We now know that DNA contains the information that is necessary for the production of proteins. The question is how the information stored in DNA can be decoded into a protein? One of the two DNA strands is transcribed into RNA. This RNA which contains the coded information, acts as a messenger molecule which is further translated into polypeptide. It's essential to understand the nature of genetic code to understand how the coded information in RNA is decoded to protein. Genetic code is a dictionary for the translation of mRNA into protein.

DNA is made up of only 4 different nucleotides (A,T,G and C) and proteins are synthesized from 20 different amino acids. The question is how 4 nucleotides could specify 20 amino acids ? A singlet code (each nucleotide codes for one amino acid) specifies only 4 amino acids, a doublet code (2 bases code for one amino acid) specifies only 16 amino acids (4^2). So the minimum number of nucleotides needed to code for 20 different amino acids is 3. This group of 3 nucleotides or nucleotide triplet is called a codon. A triplet code will contain 64 codons (4^3) which are in excess of the number of amino acids. The code was deciphered in 1960s by the important contributions made by Nirenberg, Matthaei, Gobind Khorana and Ochoa.

You may have a doubt whether the same genetic code is followed by plants, animals and bacteria as all of them have same 4 bases in their genetic material. Yes, genetic code is universal, except slightly different code is used in mitochondria and by few prokaryotes. Because of this property, we are able to translate mRNA from one species, in a cell of another species (recombinant DNA technology).

1.5.1 Properties of Genetic Code

The genetic code is triplet : The code is read in 3 letter words. A group of three nucleotides code for one amino acid.

The code is degenerate : There are 64 codons but amino acids are 20 only which means some amino acids are specified by more than one codon. Eg: GUU,GUC,GUA and GUG code for valine. All these 4 codons are said to be degenerate.

First base 5' end	Second base U		C		A		G		Third base 3' end
U	UUU	Phe	UCU	Ser	UAU	Tyr	UGU	Cys	U
	UUC		UCC		UAC		UGC		C
	UUA	Leu	UCA		UAA	Stop	UGA	Stop	A
	UUG		UCG		UAG		UGG	Trp	G
C	CUU	Leu	CCU	Pro	CAU	His	CGU	Arg	U
	CUC		CCC		CAC		CGC		C
	CUA		CCA		CAA	Gln	CGA		A
	CUG		CCG		CAG		CGG		G
A	AUU	Ile	ACU	Thr	AAU	Asn	AGU	Ser	U
	AUC		ACC		AAC		AGC		C
	AUA		ACA		AAA	Lys	AGA	Arg	A
	AUG	Start Met	ACG		AAG		AGG		G
G	GUU	Val	GCU	Ala	GAU	Asp	GGU	Gly	U
	GUC		GCC		GAC		GGC		C
	GUA		GCA		GAA	Glu	GGA		A
	GUG		GCG		GAG		GGG		G

Fig.1.7: The genetic code

The code has polarity: Codons may specify different amino acids when they are read in opposite directions.

Eg.: 5' CCU 3' → proline

3' UCC 5' → serine

As translation occurs in 5'→ 3' direction, it's apt to read the codons in 5'→3' direction only.

The code is non overlapping: Each codon consists of three consecutive nucleotides. That is none of the nucleotide is part of 2 codons. Eg. 5' GCUACCUGC 3'

Non overlapping code will specify only three amino acids. NH2-ala-thr-cys-COOH.

Overlapping code will specify seven amino acids. NH2-ala-leu-tyr-thr-pro-leu-cys-COOH.

The code is comma less: There is no gap or punctuation between 2 codons. After one amino acid is coded, the second one will be coded automatically. If there is any gap between two codons, deletion or addition of one base should not change the reading frame. But a change in reading frame was observed which means that code is comma less.

Reading Frame

The possible way in which a nucleotide sequence is read during translation is called the reading frame. Basing on the starting point, a single strand of DNA molecule can be read in three possible ways.

Eg: 5'AGCGCAAGGCGA…..3'

The above sequence has three possible reading frames –one starting with the first base, other two frames starting with second and third bases.

5'AGC GCA AGG CGA….3'

5'GCG CAA GGC GA….3'

5'CGC AAG GCG A…..3'

The code is unambiguous: Though there are few exceptions, a particular codon will always specify the same amino acid. Eg. GCU always codes for alanine.

The code contains "start" and "stop" signals: There is only one start codon (AUG) whereas termination codons (UAA, UAG and UGA) are three in number.

Wobble Hypothesis: Leaving the three termination codons which are recognised by proteins, the remaining 61 codons are recognised by tRNAs. There are only about 30 types of cytoplasmic tRNAs. Then how's it possible to interpret 61 codons? This is possible because of the relaxation of normal base pairing rules when it comes to codon-anticodon recognition. According to Wobble hypothesis of Crick, normal A-U and G-C rules are followed for the first two base positions only but wobbling occurs at third position (G can pair with C or U and U can pair with A or G).

1.6 GENE EXPRESSION

Before going to gene expression, let us first understand the meaning of gene, the number of genes in humans, their location and their structure. In simple terms, gene is a stretch of DNA that carries the information necessary for the synthesis of a polypeptide. Actually the definition of gene is much more complex. Today we know that a single gene can give rise to many polypeptides. There are about 25,000 genes in humans according to Human Genome Project. The number of proteins about two lakhs is far greater than the number of genes because of alternative splicing. Genes are located on chromosomes. Each chromosome contains many number of genes arranged in a linear order. Eukaryotic genes are split genes, which means that their coding sequence is not contiguous but it is interrupted by noncoding or intervening sequences called introns. The coding sequences or the expressed sequences are called as exons. In addition to the coding and noncoding sequences, there are flanking regions which are important in regulation and have 'start' and 'stop' signals. These include promoter which is located at the 5' end of the gene and a sequence that is present at the 3' end which provides the signal for the addition of poly A tail to the 3' end of mature mRNA.

Fig. 1.8: Structure of a gene

Gene expression is a process in which a protein is synthesized from a gene. It occurs in two major steps. The first step is transcription, in which the linear DNA is transcribed into linear mRNA. The second step is translation during which mRNA associates with the ribosomes present in the cytoplasm and directs the synthesis of proteins.

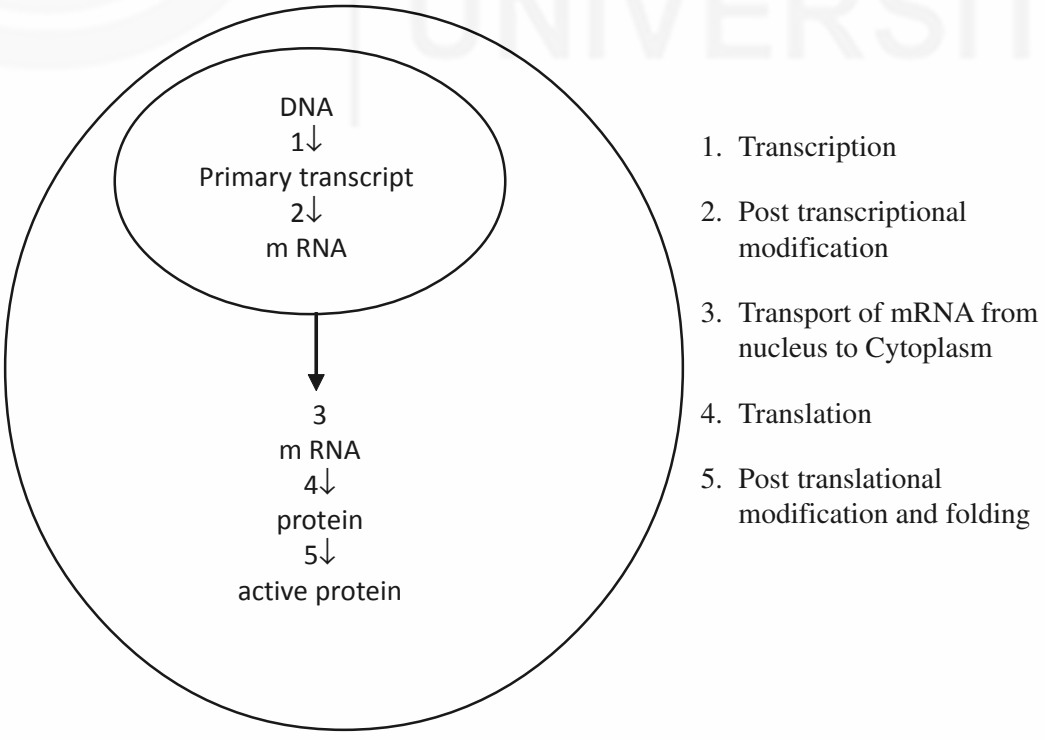


Fig.1.9: Gene expression

Here you should note the point that only a small proportion of the total DNA (1.5%) is coding. Moreover, all the genes that are transcribed are not translated that is the end product of some genes is RNA itself eg: tRNA, rRNA etc.

1.6.1 Transcription

It's a process in which single stranded RNA is generated from one of the strands of the DNA. It occurs in nucleus in 5'→ 3' direction. It needs RNA polymerase, ribonucleotides and several proteins for initiation. Only one of the two strands of the DNA acts as a template. RNA that is synthesized is complementary to the template but similar in sequence and orientation to that of nontemplate strand (except U is present in place of T). Therefore nontemplate strand is called sense strand and template strand is called antisense strand. Whenever we want to give a gene sequence, it's customary to give the sequence of sense strand in 5' to 3' direction.

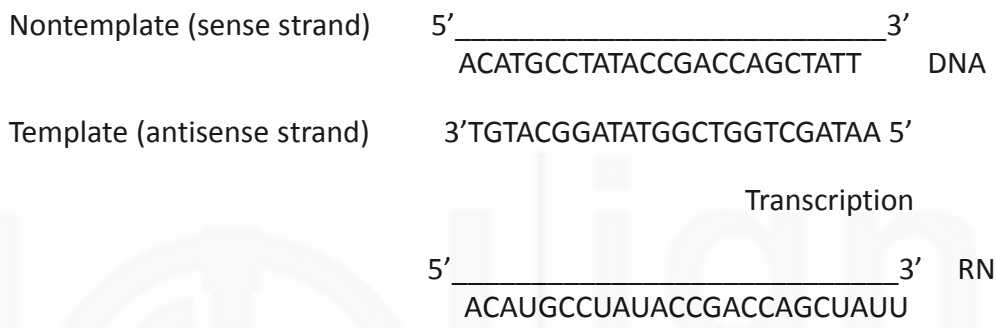


Fig. 1.10: Process of transcription showing the similarity between RNA and sense strand

The first base that's transcribed is denoted +1 and the bases that are proceeding in the right side (5'→ 3') are indicated by positive numbers and the direction is called downstream. Conversely the bases towards the left side of +1 are indicated by negative numbers and the direction is called upstream. The promoter which is present in the upstream region of the sense strand contains a group of short sequence elements called TATA box, GC box, CAAT box etc. These elements will be recognised by proteins called transcription factors (TFs). Only when TFs bind to the promoter, followed by binding of RNA polymerase, then transcription occurs. There are three kinds of RNA polymerases in humans. RNA polymerases I and III transcribe the genes which code for tRNA, rRNA and various small RNAs. Structural genes (genes which code for proteins) are transcribed by RNA polymerase II. Termination occurs in them by endonucleolytic cleavage (downstream to a sequence AAUAAA) followed by addition of poly(A) tail.

1.6.1.1 Ribonucleic Acid (RNA) and its Types

RNA molecule is also a polynucleotide chain. The sequence of RNA is determined by the DNA sequence. The difference between DNA and RNA is, it's single stranded, contains ribose sugar (2'OH) in place of deoxyribose sugar (2'-H) and thymine is replaced by Uracil. Occasionally it may fold on itself to give stem loop structures.

Transcription leads to the synthesis of several different types of RNA. Messenger RNA, ribosomal RNA and transfer RNA are the major classes of RNA involved in protein synthesis.

1.6.1.2 Messenger RNA (mRNA)

It's this RNA, which carries the message present in the gene to the cytoplasm where synthesis of protein occurs. Only the central part of the mRNA is translated. The region of the first exon and the last exon which are not translated are denoted as 5'UTR and 3'UTR. Each group of three mRNA bases constitutes a codon which specifies an amino acid. The length of different mRNAs vary considerably basing on the length of the gene.

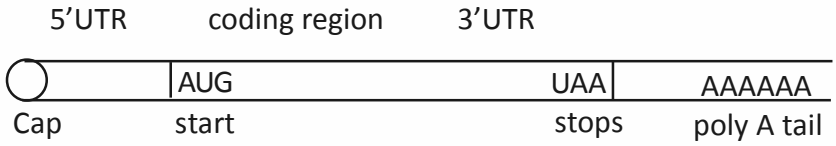


Fig.11: Structure of mRNA showing coding and untranslated regions

1.6.1.3 Transfer RNA (tRNA)

Transfer RNA molecules interpret the mRNAs with the aid of ribosomal RNAs. They are smallest RNAs (about 80 nucleotides long) which carry the amino acids to the site of protein synthesis. Their one end binds to a codon in the mRNA and the opposite end carries a specific amino acid. Thus it acts like an adaptor. Because of the formation of Hydrogen bonds between some of the complementary bases it forms a clover leafed structure (secondary structure). Its three dimensional structure is however L shaped.

1.6.1.4 Ribosomal RNA (rRNA)

About 80% of the total RNA is rRNA. Generally the largest of the RNAs, along with ribosomal proteins it forms the ribosome. A ribosome is made up of two subunits, both of which join at the time of protein synthesis.

Larger subunit: 3 kinds of rRNAs (28S, 5.8S and 5S) + about 50 ribosomal proteins. Smaller subunit: single rRNA (18S) + more than 30 ribosomal proteins

1.6.2 Post Transcriptional Modifications

All the above three types of RNAs undergo post transcriptional modifications that is; certain bases present in the RNAs are removed. The RNA that is obtained after transcription is termed primary transcript. In case of mRNA, the introns are removed and the exons are spliced together to form the mature mRNA. Splicing occurs with the help of certain conserved sequences present in the introns. In addition to splicing , in case of mRNA, a cap and a poly A tail are added to the 5' and 3' ends of the mRNA respectively. These two structures help in the migration of the mRNA from the nucleus to the cytoplasm and also in the regulation of gene expression.

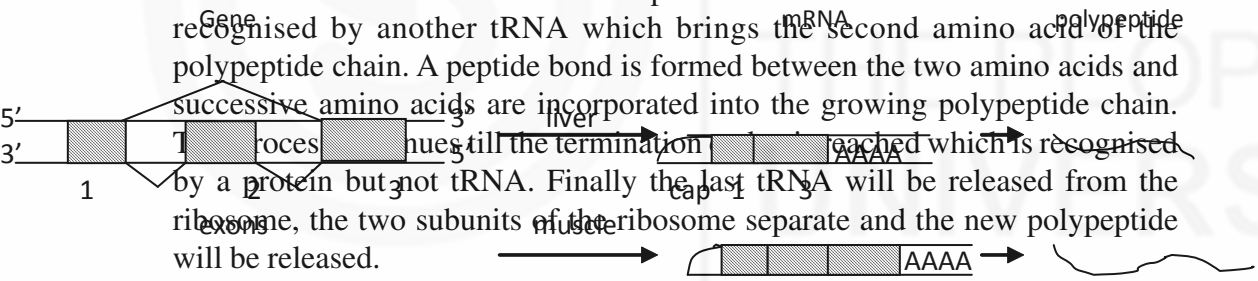
We've already discussed about alternative splicing due to which we are getting about 2,00,000 proteins from the 25,000 genes we have. In simple terms, alternative splicing means getting more number of mRNAs from a single gene. This is possible by differential splicing in different tissues. For eg. a sequence which acts as an intron in one tissue may act as a coding sequence in another tissue thereby changing the sequence of the polypeptide in different tissues.

Fig. 1.12: Alternative splicing where liver is using only two exons whereas all the three are used in the muscle

1.6.3 Translation

It's a process in which the information present in an mRNA is decoded into the amino acid sequence of a protein. It requires mRNA, tRNA, ribosomes, ATP and various protein factors. It occurs on ribosomes in the cytoplasm. It also occurs in 5'→3' direction. The 5' end of mRNA corresponds to the amino terminus of the protein. Translation starts from the initiation codon and ends with the termination codon. That's the reason why most of the polypeptides start with methionine.

Initiation of translation requires several factors which include a cap binding protein, initiation factors, smaller subunit of the ribosome, initiator methionyl tRNA, all of which bind to the 5' cap region of the mRNA. The initiation complex formed scans the mRNA for the initiation codon. In the elongation step, larger subunit attaches to the initiation complex. The codon next to the AUG is then recognised by another tRNA which brings the second amino acid of the polypeptide chain. A peptide bond is formed between the two amino acids and successive amino acids are incorporated into the growing polypeptide chain. The process continues till the termination codon is reached which is recognised by a protein but not tRNA. Finally the last tRNA will be released from the ribosome, the two subunits of the ribosome separate and the new polypeptide will be released.



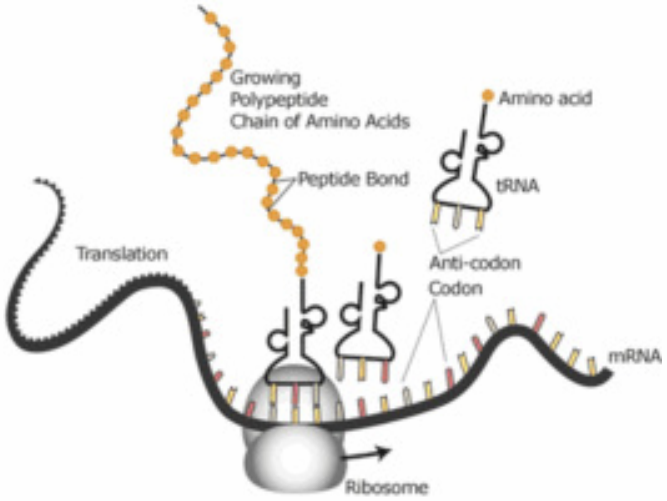


Image adapted from: National Human Genome Research Institute.

Fig.1.13: Translation of mRNA into polypeptide (Source:http://www2.le.ac.uk/departments/genetics/vgec/diagrams/47-translation.gif)

1.6.3.1 Proteins

A protein is a polymer made up of amino acids. It's the end product of most of the genes. Proteins perform all the metabolic reactions that are carried out in a cell. Though the term protein and polypeptide are loosely used, there is a difference between the two. Polypeptide is the molecule that's formed after translation. After its release, the nascent polypeptide folds up and achieves a three dimensional conformation to become functional protein. Many proteins depend on other proteins called chaperones for folding. In addition to proper folding, polypeptides also undergo post translational modifications (hydroxylation, glycosylation, phosphorylation etc.) to achieve functional status. So a polypeptide is a precursor of protein. Some proteins may have more than one polypeptide, which may be of same kind or of different kinds.

1.7 REGULATION OF GENE EXPRESSION

There are 200 different types of cells in human beings. All of them have the same DNA content or to be precise all the genes. Yet all of them are morphologically different and have different function. This depends on the type of genes that are expressed in these cells or in other words differential gene expression is responsible for the diverse properties of different cells. All the genes are not expressed in all the cells or all the times. By this we mean that some genes are expressed only in some cells (tissue specific expression). Similarly some genes are expressed only during a particular time of development. However there are certain genes which are expressed in all cell types (house keeping genes) eg: genes for rRNA, tRNA, DNA polymerases etc. The question is how does a cell know which genes to express, when to express and to what extent. That's what is explained in this topic.

It's a waste of energy for the cell to produce the proteins which are not needed by it. Cells have their own methods, by which they can regulate the expression of genes. Interestingly it's the proteins which are largely responsible for regulation of gene expression. Regulation occurs at three levels.

1) Transcription is the predominant stage at which regulation occurs. Transcription occurs at basal level with the help of TFs. Up regulation and down regulation of gene expression is possible with the help of proteins called activators and repressors respectively. Activators bind to sequences called enhancers whereas repressors bind to silencers. These sequences may be located near the promoter region or far away in the upstream or downstream region of the gene. These regulatory proteins are controlled by signals which determine whether these proteins bind DNA. They determine the amount of the protein to be synthesized. Tissue specific expression is possible by limiting the availability of TFs needed by a gene only to a particular tissue. In some cases, the promoter sequence is methylated in all other tissues except the tissue where it's expressed. Histone proteins also have a role in regulation. Methylation of certain amino acids in histone proteins turns off the expression of a gene.

2) Regulation also occurs at post transcriptional level. Alternative splicing produces different isoforms in different tissues. Isoforms also result due to alternative polyadenylation (same gene uses different polyadenylation signals

in different tissues) and RNA editing (same gene produces different isoforms due to single base substitution, deletion or insertion at the RNA level) .

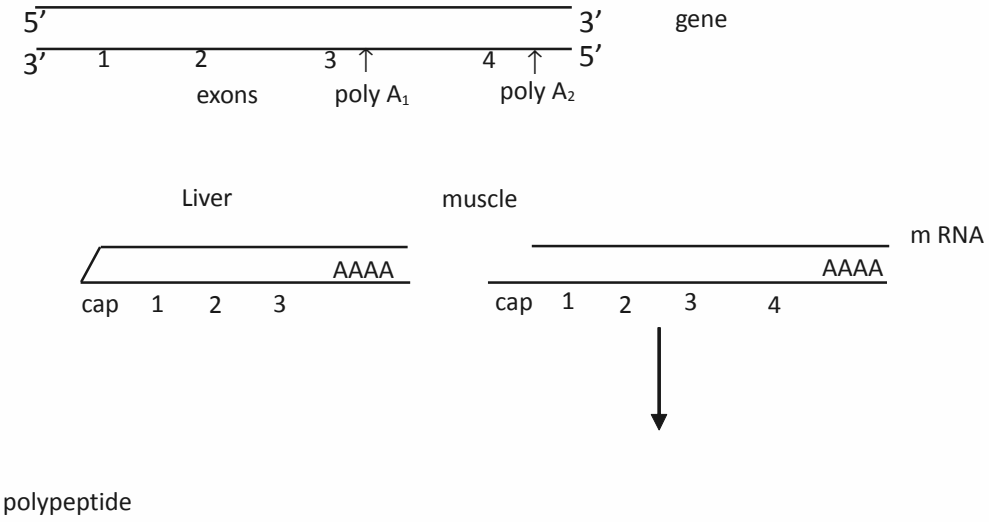


Fig. 1.14: Alternative polyadenylation where liver uses polyA1 signal and muscle usespolyA2 signal

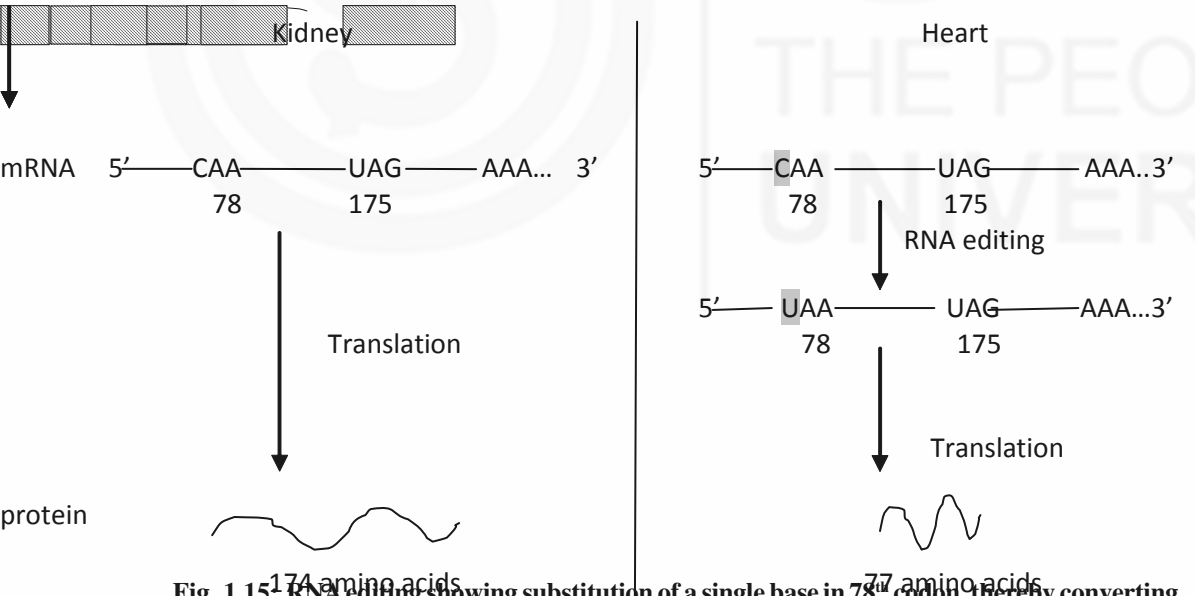


Fig. 1.15: RNA editing showing substitution of a single base in 78th codon, thereby converting it into a stop codon in heart. Thus, a single gene expresses as 174 amino acid protein in kidney and 77 amino acid protein in heart.

3) Regulation at translation level includes longevity of mRNA which depends on length of poly A tail (mRNAs without poly A tail are short lived), structure of 3'UTR (many repeats of AUUUA in the 3'UTR makes the RNA short lived) etc. Translation of some mRNAs is regulated by specific RNA binding proteins. Degradation of mRNA is another control point. Rapid degradation of mRNA prevents undesired protein synthesis.

1.8 SUMMARY

Molecular Genetics deals with the study of gene expression and its regulation. Revolutionary changes occurred in the field of molecular genetics due to the invention of molecular techniques. It has several applications like diagnosis of genetic as well as infectious diseases, treatment of the disease, personal identification etc.

DNA is the information macromolecule. Each chromosome contains a single DNA molecule. The structure of DNA was deciphered by Watson and Crick. DNA is a right handed double helix with two ribbon like strands constituting the sugar phosphate backbones. The horizontal rungs are made up of nitrogenous bases. The two strands are complementary and antiparallel to each other. They are held together by Hydrogen bonds formed between the opposite bases. The order of the bases in the DNA determines the hereditary features.

The human genome is comprised of two genomes : nuclear genome which is complex and comprises the bulk of the genome and mitochondrial genome. The size of the haploid genome is about three billion bp and it is distributed in 23 chromosomes. The organisation of the genome is complex. Neither the distribution of the genes nor the base composition is uniform throughout. There is high degree of variation not only in the size of the gene but also in the size of exons and introns. In addition, the existence of repetitive sequences, transposons, pseudogenes, overlapping genes and genes within genes make the genome more complex.

The size of the mitochondrial genome is 16,569 bps. It is a circular, double stranded molecule. It shows maternal inheritance. It contains 37 genes. Most of the genes needed for its function are coded by nuclear genome. Mutations do occur in mitochondria which leads to disease.

Genetic code is the relationship between the nucleotide sequence of mRNA and amino acid sequence of protein. The genetic code is triplet, degenerate, nonoverlapping, commaless and universal.

Gene expression is a two step process. First the information present in DNA is transcribed to RNA. Later the information in the mRNA form is decoded to amino acid sequence of the polypeptide by translation. The polypeptide that is formed during translation, undergoes folding and post translational modifications to form a functional protein.

Gene expression is highly regulated. Each cell contains all the genes present in the total DNA. However to save energy, cells express only the proteins needed by it, in required amounts and at needed time. This kind of control over gene expression is achieved largely by proteins. Gene expression is regulated at different levels like transcription, post transcription, translation. Major control occurs at the transcriptional level. Same gene can produce different forms of proteins in different tissues by mechanisms like alternative splicing, alternative polyadenylation and RNA editing.

Suggested Reading

Strachan, T and Read, A.P. 2004. *Human Molecular Genetics*. Wiley-Liss.

Lewin, B. 2004. *Genes VIII*. Prentice Hall.

Watson, J.D, Baker, T.A, Bell, S.P, Gann, A, Levine, M and Losick, R. 2004. *Molecular Biology of the Gene.* Pearson Education.

McConkey, E. H. 1993. *Human Genetics. The Molecular Revolution*. Jones and Barlett Publishers.

Sample Questions

1) Describe the salient features of DNA double helix.

2) "The organisation of human nuclear genome is complex". Justify the statement.

3) Define genetic code and write about its properties.

4) Explain how the nucleotide sequence present in a gene is used to synthesize a protein.

5) Describe how gene expression is regulated at different levels.

Short Notes

i) MtDNA

ii) Genome

iii) RNA and its Types

UNIT 2 DNA POLYMORPHISMS

Contents

Learning Objectives

After having studied this unit, you will be able to:

- understand Human DNA polymorphisms;
- explain types of DNA polymorphisms;
- discuss DNA polymorphisms role in disease manifestation; and
- describe Single Nucleotide Polymorphism's role in reconstructing Human evolution and modern Human migrations.

2.1 INTRODUCTION

Human genome (entire genetic material present in a cell) consists of 3 billion bases. There are about 10 million Single Nucleotide Polymorphisms. The DNA sequence will carry code of information for carrying genetic information from parent to child (generation to generation). A person or plant or animal phenotypically looking different, means that there are variations in the genetic material of the organism. Any change in the DNA sequence will bring change in the genetic information, inturn brings change in phenotypic expression and biological function. The change in the DNA sequence is called *Mutation*. If the mutation frequency is more than 2 per cent in a population, it is called as *polymorphism.* Hence, we can define DNA polymorphism as DNA having more than one form, with a frequency of above 2 percent in a population.

2.2 DIFFERENT FORMS OF DNA POLYMORPHISMS

DNA polymorphisms can be studied in the form of Single Nucleotide Polymorphisms (SNPs), Restriction site Polymorphisms (RSPs) or Restricted Fragment Length Polymorphisms (RFLP) and Variable Number of Tandem Repeats (VNTRs).

Single Nucleotide Polymorphisms (SNPs): A single nucleotide is substituted by a different nucleotide.

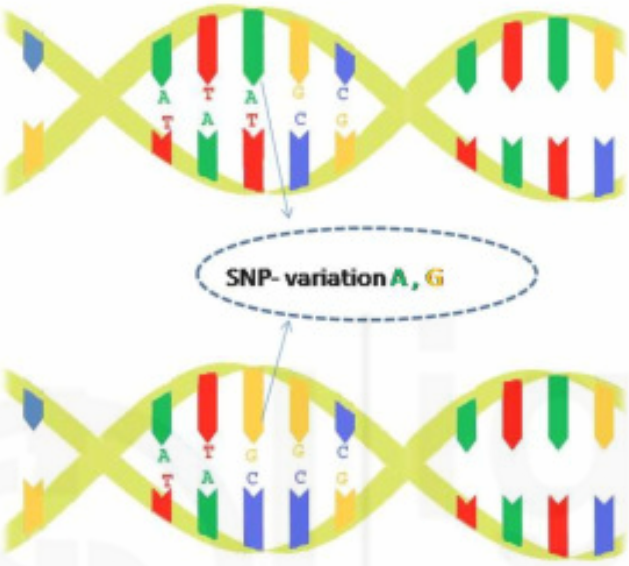


Fig.2.1: SNP in double strand DNA

There are two types of nucleotide substitutions resulting SNPs. 1) Transition: A substitution occurs between Purines (A,G) or between Pyrimidines (C,T). This type of substitution constitutes two thirds of all SNPs. 2) Transvertion: A substitution occurs between a Purine and a Pyrimidine.

Insertions: A new nucleotide will be inserted in the sequence.

Deletions: An existing nucleotide will be deleted in the sequence.

Restriction site Polymorphisms (RSPs) or Restricted Fragment Length Polymorphisms (RFLP): A sub set of SNPs cause a loss or gain of a restriction site (restriction site is the location where a particular enzyme cuts the DNA sequence at particular sequence location into pieces of DNA). Due to change in nucleotide at particular site location, enables the enzyme to cut DNA into pieces. This leads to create different length of DNA piece in an individual and another length of DNA piece in another individual. This is called Restriction site polymorphism. RSPs are described as restricted fragment length polymorphisms (RFLP).

III.Variable Number of Tandem Repeats (VNTRs): It is divided into two types:

Microsatellite polymorphism

Minisatellite polymorphism.

Micro-satellite polymorphism is also called as Short Tandem Repeats (STRs). A small array of tandem repeats of a simple sequence (usually less than 10 base

pairs). Ex; GATAA GATAA GATAA GATAA GATAA GATAA in this sample 5 bases repeated 6 times.

a) GA GA GA GA GA – it is dinuleotide repeat

b) TAT TAT TAT TAT- it is trinuleotide repeat

Mini-satellite polymorphism: A collection of moderately sized arrays of tandemly repeated DNA sequence which are dispersed over considerable portions of the nuclear genome.

c) TTAGGGTACCGG TTAGGGTACCGG TTAGGGTACCGG –this array of 12 nucleotides repeats from 3-20 Kbp(thousand base pairs).

2.3 HUMAN EVOLUTION WITH SPECIAL REFERENCE TO MITOCHONDRIAL DNA AND Y-CHROMOSOME POLYMORPHISMS

DNA polymorphisms particularly SNPs became a powerful tool in reconstructing human origins, evolution and prehistoric migrations. Earlier we used to depend on archaeological and paleontological evidences to reconstruct human evolution. In the absence of these traditional evidences DNA analysis became an alternative tool to reconstruct human evolution.

The hominid fossil record in Africa begins about 4 MYs ago in Early Pliocene with representatives of the genus Australopithecus from Ethiopia and Tanzania. Homo erectus arose more than a million years ago in the Pleistocene, giving rise to our own genus, Homo. Anatomically modern humans began to appear 120000-100000 years ago and co-existed with Neanderthals until the latter became extinct about 30,000years ago. Based on these evidences, two theories have been proposed for the evolution of modern humans: **1.Multiregional evolution**: It proposes that present day worldwide populations are the descendants' of in situ evolution after an initial dispersal of Homo erectus from Africa during the Lower Pleistocene(~650kybp), **2. Uni regional hypothesis** (also called as **Recent African Origin model** or **Out-of-Africa**):- All present day populations have descended from a recent common ancestor that lived in East Africa ~150,000 years ago.

At this juncture, mitochondrial DNA and Non-recombining region of Y-chromosome (NRY) analysis provided an alternative approach to reconstruct Modern Human evolution. mtDNA and NRY Y-chromosome analysis enable us to trace maternal and paternal lineages of modern humans. Along with these DNA markers autosomal and X-linked markers have also been studied.

DNA can also be extracted from bone material of ancient specimens. The ancient mtDNA analysis from Neanderthal specimens reveals that Neanderthals are not immediate ancestors to modern humans. Modern humans diverged from Neanderthals about 400,000 years ago. Neanderthals went extinct without contributing any mtDNA to modern humans.

2.3.1 mtDNA Polymorphism-Human Evolution

DNA polymorphisms suggested a recent origin of modern humans from African populations. Initial evidence came from mtDNA, which is transmited maternally. Each human cell cytoplasm contains 10-100 mitochondria. Each mitochondrion will have a circular double strand DNA molecule about 16569 base pair length. The most ancient mtDNA haplotypes (having the same genotype) are L0, L1, L2

and L3. Haplotypes L1 and L2 are specific to the sub-Saharan Africa. L3 is present in North East Africa and Middle East. L3, M & N are parallel branches. M branch is called as Asian branch and N is called as European branch. All branches of M arose in Asia. M branches didn't present in Europe. M1, a branch of M present in East Africa is originated in Middle East and back migrated into Africa. In Asia N branch is also present. M is the oldest branch than N branch. An ancestral branch of Asian might have arisen in North East Africa and subsequently left to colonise Asia (50-70 thousand years ago) and Europe (45-50 thousand years ago).The mtDNA analysis of world populations reveals that modern humans can be traced back to a single mother, 'mitochondrial eve'. The analysis also shows that this individual existed about 100,000-130,000 years ago in east Africa. Of course, the mitochondrial eve was not the only person living on the planet at that time: there are perhaps about 10,000 individuals living at that time but unlike Eve, there mtDNA sequences didn't get transmitted to the present human populations.

SNPs normally exhibit two forms of variation at a nucleotide position or have two alleles. For example Africans (L0, L1 & L2 haplogroups) have Thymine at nucleotide position 3594 where as all Non-Africans will have Cytocine at np3594. Likewise all Asians have four mutations (M lineage) in their mtDNA sequence.

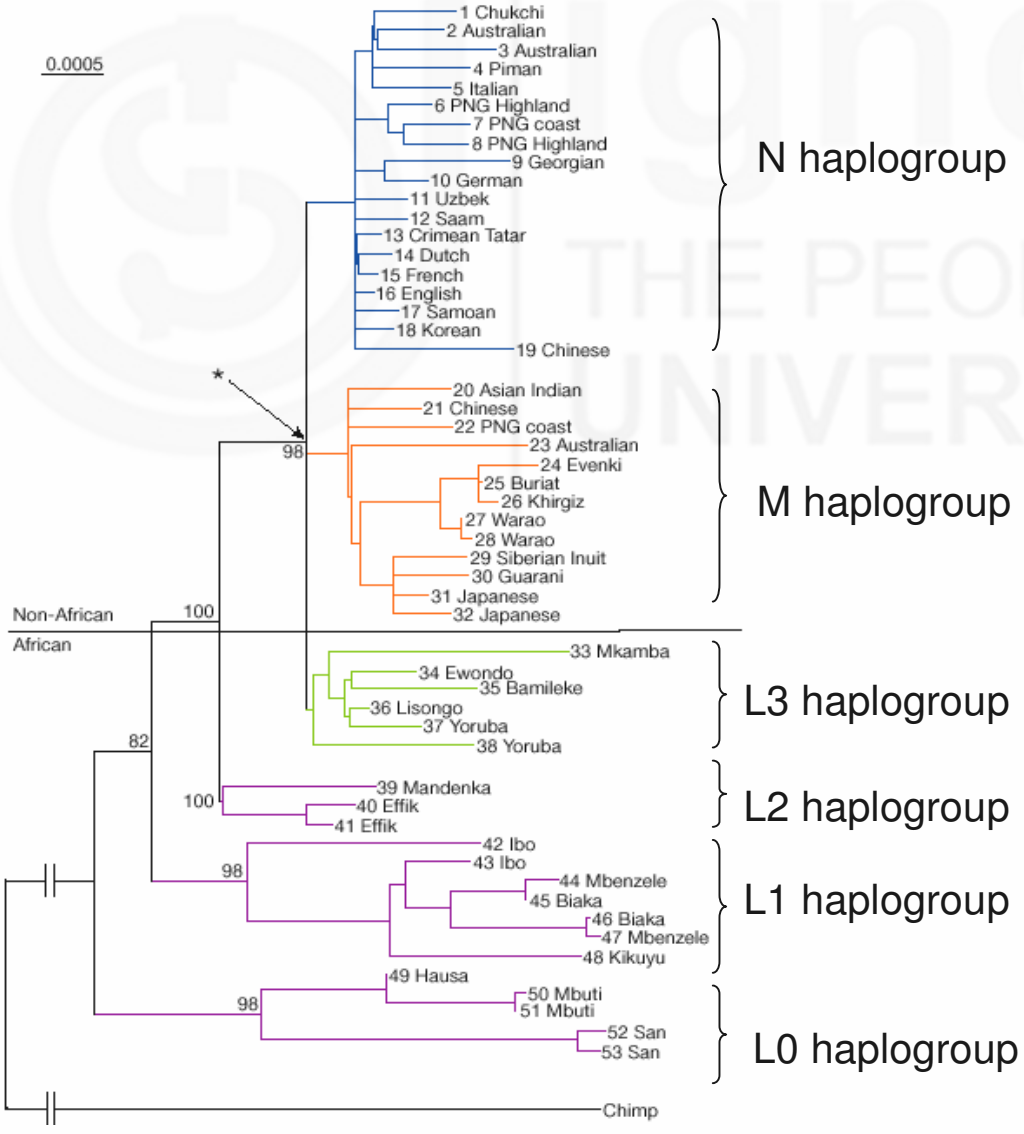


Fig.2.2: mtDNA phylogenetic tree constructed based on Polymorphisms present in mtDNA of world populations (adopted after Ingman, 2000)

Indian scenario

The initial dispersal of modern humans from East Africa en-route North and East of Africa has now been documented, following the African mtDNA haplogroups into Saudi Arabia and then Western India. Indian specific mtDNA branches, M and N encompass all the populations in India irrespective of their social rank, caste or tribe. In India the frequency of M haplogroup ranges from 54 to 97 percent. Indian specific M sub haplogroups are M2,-M6, M18, M25, and M30-M62. M1is present in North East Africa, M31 & M32 are specific for Andaman Islands. M7, M8, M9, M10, M11 & M12 are specific for China and Japan. Most numerous sub haplogroups of European N haplogroup are Indian specific. Ex: N5, R5, R6, R7, R8 T30, R31, U2. Genetic links of Indians with East Eurasians, West Eurasians and Australians are established by mtDNA polymorphisms.

2.3.2 Y Chromosome Polymorphism-Human Evolution

The Y chromosome is a suitable tool for investigating the recent human evolution, for medical genetics, DNA forensics and genealogical reconstructions, due to its uniqueness among the other human chromosomes. The Y chromosome has a sex-determining role, it is male specific and constitutively haploid (Single). It is inherited paternally and is transmitted from father to son, and unlike other chromosomes, the Y chromosome escapes meiotic recombination in its NRY (Non Recombining Y chromosome) region. The non-recombining portion of the Y chromosome descends as a single locus. As they change only by accumulating mutations in time, they preserve by far more simple record of their history compared to autosomes.

Y chromosome variation consists of large amount of different types of polymorphisms, which are widely used in evolutionary studies. They may roughly be divided into two large groups: bi-allelic markers and polymorphisms of tandem repeats or multi-allelic markers. Bi-allelic markers include SNPs (Single Nucleotide Polymorphisms) and insertions and deletions (indels). SNPs are the most common type of polymorphisms, constituting more than 90% of total polymorphisms of DNA. Only these bi-allelic mutations that have occurred, only once in history of humans and have a detectable frequency in human populations are used in phylogenetic studies.

Y chromosome DNA polymorphisms are useful to trace paternal lineages. The Y chromosome consortium is formed to document the binary polymorphisms in NRY (non recombining region of Y chromosome). There are about 599 polymorphisms made Y chromosomes in to 311 groups. This information was constructed into a tree and named the main branches starting from alphabets A to T. Major branches are called as Clades and sub branches are called as a haplogroups.

Polymorphisms at P91, M168, M294 distinguishes A, B clades from the rest of clades. Clade A & B are exclusively present among African populations. The majority of branches of the Y chromosome tree outside Africa are composed of a tripartite assemblage of the following haplogroups: a) C; b) D and E, and c) an overarching haplogroup F that defines the internal node of all remaining haplogroups from G to T.

Because the mutation defining haplogroup C (M130=RPS4Y) has not been observed in any African populations, this haplogroup is likely to have arisen somewhere in Asia after an early departure of modern humans from Africa, prior to the arrival of them to Sahul in Southeast Asia. The most western region where

haplogroup C* has been detected is India. This lineage consists of several sub-lineages with irregular phylogeographic patterning, ranging from Central and North Asia to America and in the direction of Southeast Asia up to Australia and Oceania. Differently from hg C, haplogroups E and D share three phylo-genetically equivalent markers.

Calde D* is found in Andaman Islands whereas D1 & D2 are found in Tibetans and Japanese. E is the most frequent and divergent in Africa.The third major sub-clade of M168 lineages is super haplogroup F. It is characterized by mutation M89 at its root from which all other haplogroups deploy. F has been suggested to have evolved early in the diversification and migration of modern humans. Later on, the ancestral trunk of F diversified into many branches by subsequent acquisition of mutations, giving rise to many region-specific haplogroups, such as J and G in Near and Middle East, I in Europe, H in Southern Asia, etc. An expansion of F lineages gave rise also to a population that acquired the M9 mutation (haplogroup K), which defines another major bifurcation in the phylogeny . The branches of this clade probably migrated in different directions (North and East) and gave start to many separate and region-specific haplogroups in Eurasian continent and beyond. Out of descendants of M9 lineage, haplogroup L (M20) has greatest frequency in Southwest Asia and distinctive K lineages and M (M4, M5) haplogroup are restricted to Oceania and New Guinea, whereas haplogroup O with its numerous sub-clades predominates in southern and southeastern Asia, reaching North China, Manchuria and some Siberian populations. The population carrying M9 expanded also in direction of north towards Central Asia characterized by subsequent mutations defining haplogroup P, which encompasses distinctive eastward expanding haplogroup Q (M242) characteristic to Siberian populations and Amerindians and Eurasian haplogroup R lineages that have expanded westward. Thus, one may speculate that multiple independent formations and fragmentations of populations carrying F-related lineages throughout most of Eurasia may have displaced the earlier haplogroup C and D lineages towards the margin in many areas. Among Indians H, O, R1, R2 are the major clades. H haplogroup is nearly restricted to India, Srilanka and Pakistan. Among Austro-Asiatic language speaking groups of India O haplogroup is predominant followed by H group. Indo-European language speakers have 50 percent of O haplogroup followed by H & R.

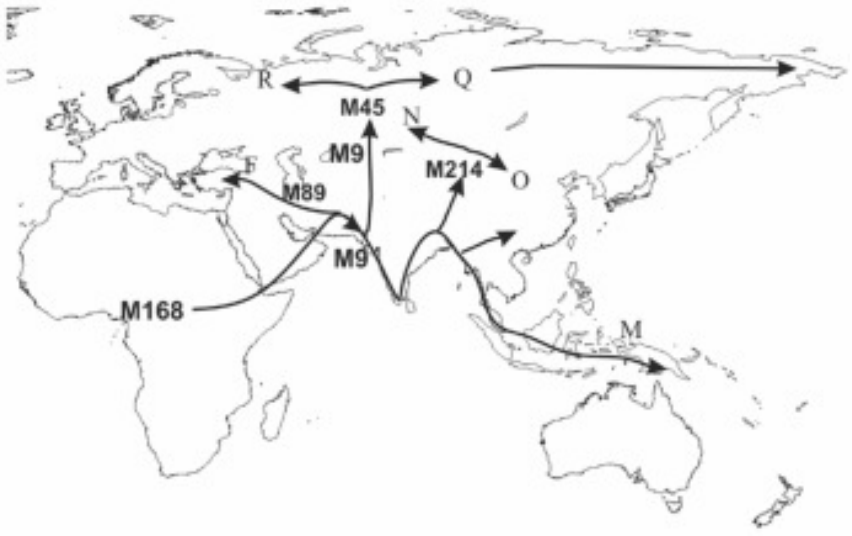


Fig.2.3: Schematic reconstruction of super haplogroup F (defined by M89) origin, subsequent diversification of M9 lineages and their possible migration routes across the world. Adapted from Underhill (2003).

Table 2.1: Number of mutations associated with 20 major Y chromosome clades (Karafet,2008)

Clade	Mutations	haplogroups
A	47	12
BT	5	
B	32	17
CT	3	
C	30	19
DE	8	
D	23	15
E	101	56
C,FT	1	
FT	25	
F	6	5
G	13	10
H	12	10
IJ	7	
I	35	16
J	42	34
KT	4	
K	5	5
L	13	7
M	20	12
NO	6	
N	11	10
O	48	31
P	20	1
Q	18	14
R	50	28
S	8	6
T	6	3
TOTAL	**599**	**311**

Based on the above table the tree has been constructed which is as follows.

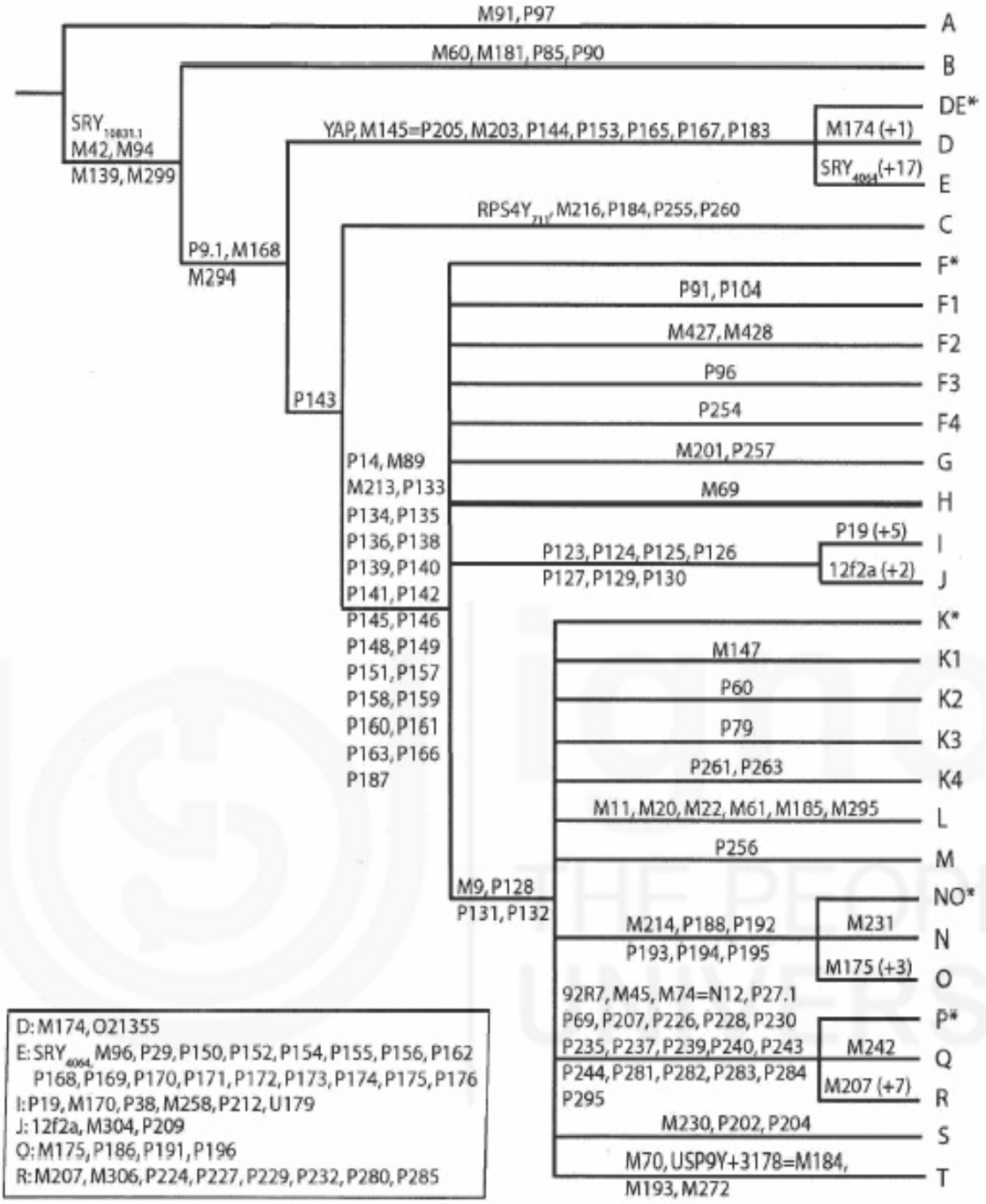


Fig.2.4: Y-Chromosome tree. Mutation names are indicated on the branches. A to T are branch names. Sub trees are not shown in this figure (Karafet, 2008)

Apart from bi-allelic polymorphisms, insertions and deletions (indels) persist over generations and are sufficiently common to be considered as polymorphisms. One such example is a 2kb deletion in 12f2 marker, used for defining haplogroup J. Some indels have arisen independently more than once in human history. For example, the deletion or duplication of the 50f2/C region in background of different haplogroups is thought to be arisen at least 7–8 times. Another example is the deletion of DAZ3/DAZ4 region that has been indicated to occur in haplogroup N individuals, widely spread in northern Eurasia.

Another frequent type of polymorphism, present also in Y chromosome, is tandem repeats, mostly in non-coding DNA regions. According to their length, these repeats are classified as satellite-DNAs (repeat lengths of one to several thousand base pairs), mini-satellites or variable number of tandem repeats (VNTRs), ranging from 10 to 100 bp, and microsatellites or short tandem repeats (STRs), with motifs less than 10 bp, mostly 2 to 6 bp long. In Y-chromosomal studies microsatellites are widely used, than mini-satellites. Microsatellites are multi-allelic markers with different allele numbers ranging from 3 to 49 in locus. Their mutation rate is much higher than that for biallelic markers and, therefore, they are widely used in phylogenetic studies to investigate details of demographic events that have occurred in a more recent time scale. In evolutionary studies STRs are valuable in combination with binary haplogroup data, as they enable us to study diversity within a haplogroup. STRs are particularly widely explored in forensic work. So far the number of widely used Y chromosomal STRs has been quite low (about 30) but in a recent study 166 new and potentially useful STRs were described.

Based on the Phylogenetic analysis it was concluded that, all humans have originated from an African ancestor. About 70,000 thousand years ago modern man came out of Africa and peopled all the continents.

Thus the combination of molecular age and geographical structure makes mtDNA and the NRY a sensitive genetic index capable of tracing the micro evolutionary patterns of noval modren human diversity. Mitochondrial DNA and Y chromosome studies in Indian populations reveals affiliation with Europeans, East Asians, Austro-Melanesians and in situ development of deep rooted ancestry whose relative clustering and coalescence ages suggest shaping of Indian gene pool during late pleistocene.

2.4 DNA POLYMORPHISMS AND DISEASE ASSOCIATION

Disease is a disordered or incorrectly functioning of organ, part, structure, or system of the body resulting from the effect of genetic or developmental errors, infection, poisons, nutritional deficiency or imbalance, toxicity, or unfavourable environmental factors.

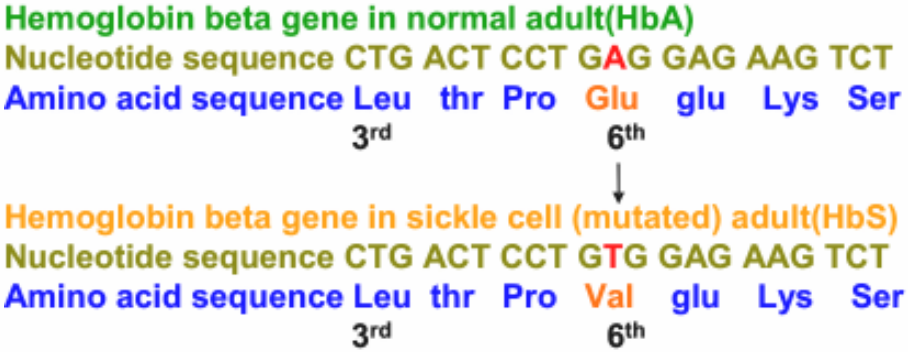
A genetic disease is any disease that is caused by an abnormality in an individual's genome. The abnormality can range from minuscule to major from a discrete mutation in a single base in the DNA of a single gene to a gross chromosome abnormality involving the addition or subtraction of an entire chromosome or set of chromosomes. Some genetic disorders are inherited from the parents, while other genetic diseases are caused by acquired changes or mutations in a pre-existing gene or group of genes. Mutations occur either randomly or due to some environmental exposure. There are about 30,000 genes in humans. Some of the mutations in these genes cause a disease, predispose us to the common diseases in combination with other variants and with the environment. Knowledge of these polymorphisms offers tremendous advantage in the study of disease and variable response to treatment.

2.4.1 Monogenetic Disease

It is caused by changes or mutations that occur in the DNA sequence of a single gene, also called Mendelian disorder. There are more than 6,000 known single-gene disorders, which occur in about 1 out of every 200 births. Some examples of monogenetic disorders include: Cystic Fibrosis, Sickle Cell Anaemia, Marfan syndrome, Huntington's disease, and Hemochromatosis. Single-gene disorders are inherited in recognisable patterns: autosomal dominant, autosomal recessive, and X-linked.

Example: Sickle cell anaemia is a disease passed down through families in which red blood cells form an abnormal crescent shape (Red blood cells are normally shaped like a disc.) Sickle cell anaemia is caused by an abnormal type of haemoglobin called haemoglobin S. Haemoglobin is a protein inside red blood cells that carries oxygen. Haemoglobin S changes the shape of red blood cells, especially when the cells are exposed to low oxygen levels. Then the red blood cells become crescent shaped or sickles. The sickling occurs because of a mutation in the haemoglobin gene. The haemoglobin beta(HBB) gene is found in region 15.5 on the short (p) arm of human chromosome 11. In sickle cell haemoglobin (HbS) the glutamic acid in position 6 is mutated to valine in a beta chain. This change allows the deoxygenated form of the haemoglobin to stick to itself and become crescent shape.

Fig.2.5: Normal and sickle red blood cells.



The fragile, sickle shaped cells deliver less oxygen to the body's tissues. They can also get stuck more easily in small blood vessels, and break into pieces that interrupt healthy blood flow.

Sickle cell anaemia is inherited from both parents. If you inherit the haemoglobin S gene from one parent and normal haemoglobin (A) from your other parent, you will have sickle cell trait. People with sickle cell trait do not have the symptoms of sickle cell anaemia. The children of both sickle cell parents will get sickle cell anaemia.

Sickle cell disease is much more common in people of African and Mediterranean descent. It is also seen in people from South and Central America, the Caribbean, and the Middle East.

2.4.2 Multifactorial Disease

It is called complex or polygenic disease. Complex diseases are caused by a interaction of environmental factors and mutations in multiple genes. Some common chronic diseases are multifactorial in nature. Examples of complex diseases include: Cardio Vascular diseases, high blood pressure, Alzheimer's disease, arthritis, diabetes, cancer, and obesity. For example, different genes that influence breast cancer susceptibility have been found on chromosomes 6, 11, 13, 14, 15, 17, and 22.

Mutations in BRCA1 gene (BRCA1 Gene is located on chromosome17q21.31) contribute significantly to the development of familial/hereditary breast and ovarian cancer. Founder mutations such as the BRCA1-185delAG and 5382insC are found among Ashkenazi Jews.

Polymorphism in BRCA1 `Chr17 at np` 37043496 is shown in the figure.

`CCGCCCCTACCCCCCC`**`G`**`TCAAAGAATACCCAT(normal form)`

`CCGCCCCTACCCCCCC`**`C`**`TCAAAGAATACCCAT (mutated form)`

Large rearrangements, mostly deletions in regions of Y-specific genes (AZFa, AZFb, AZFc), have been known as causes for many diseases leading to male infertility, causing spermatogenic failure, azoospermia, severe oligo spermia or otherwise severely impair male reproductive fitness.

2.5 TECHNIQUES IN MOLECULAR GENETICS

To study DNA polymorphisms initial step is extraction of DNA from cells. We can extract up to 1500 nano grams of DNA from 5ml of blood by Phenol chloroform method.

DNA Extraction- Principle

The process of DNA extraction can be divided into three stages: (i) disruption of the cellular membranes, resulting in cell lysis, (ii) protein denaturation, and finally (iii) the separation of DNA from the denatured protein and other cellular components.

About 5 ml of blood is transferred to a sterile 15ml conical bottom polypropylene tube and equal amount of RBC lysis buffer is added. Shake the tube gently for 3 to 5 minutes. The contents of the tube transforms into transparent red colour upon the lysis of RBC. We have to ensure that RBC is completely lysed before proceeding to further step. Once RBC is lysed, transfer the tubes into the centrifuge. Run the centrifuge at 1500 rpm for 15 mts at 20^{o}C. Upon the completion of the run carefully, remove the tubes from the centrifuge and decant the supernatant without disturbing the pellet at the bottom. Now add 4 ml of RBC buffer to the tube and break the pellet using hand or vortex. Repeat the centrifugation step as earlier with same settings. Decant the supernatant, if pellet is still red repeat the RBC lysis. If pellets are light pink or white proceed for further step.

Add 1 ml of digestion buffer and 10 ul of Proteinase K to the tube and carefully dislodge the pellet from the bottom of the tube. Adjust the hot water bath at 55^{o}C, now transfer the tubes into the hot water bath. Gently dislodge the tubes for every 30 minutes to enhance the digestion process. The content of the tubes turns clear and transparent upon the digestion.

Now add 250ul of 5M sodium per chlorate to the tube and gently mix the contents by partially inverting the tube. Now add 500ul of tris saturated phenol, 500ul of 24:1 chloroform isoamyl alcohol, mix the contents thoroughly and adjust the centrifuge at 4ºC, 4000 rpm and 15 minutes. Take a fresh tube, carefully transfer the supernatant using the 1ml pipette and cut tips into the fresh tube. Now add 500ul of 24:1 chloroform and isoamyl alcohol, and repeat the centrifugation step with same settings. Carefully transfer the supernatant using a cut tip into a fresh tube. Add double the volume of chilled alcohol; gently invert the tube for a minute. A milky white fibrous DNA is visible. Now transfer the DNA into a 1.5ml tube, add 1 ml of 70 % alcohol, spin at 12,000 rpm for 10 minutes, and repeat the step for one more time to eliminate the remaining protein contamination. Dry the pellets and add 200ul of TE buffer and mix the contents thoroughly and transfer into the hot water bath/dry bath for digesting the DNA. This process usually takes 2 hours. Transfer tubes for -80ºC for long term storage.

Details of the buffers and reagents used for DNA extraction are given below.

1) *RBC Lysis Buffer contains* Sucrose, 1M Magnesium Chloride 1M Tris-Hcl and Triton X.

2) *Digestion Buffer constitutes* 1M Tris-Hcl (pH 8.0), 1M Sodium Chloride, 0.5M EDTA (Na salt)

2) *Tris-EDTA Buffer made up of* 1M Tris-HCl (pH 8.0), 0.5M EDTA

After DNA was completely dissolved in the TE buffer, its quantity and quality was checked by both spectrophotometry and gel electrophoresis.

Determination of DNA concentration by Spectrophotometry

Prior to any analysis, DNA should be quantified and checked for purity and integrity. Based on its structure, DNA absorbs light in the ultraviolet range,

specifically at a wavelength of 260nm. A value of 1 at OD_{260} is equal to 50ng/µl double stranded DNA. Therefore to calculate the concentration of DNA, the following formula can be used:

$$\text{Concentration of DNA} = 260nm_{abs} \times 50ng/\mu l$$

Procedure

2µl DNA sample was diluted to 200 µl with Double Distilled water (Dilution 1:100). Spectrophotometer was set to auto zero with the Double Distilled water. Optical Density (OD) of the diluted DNA aliquot was measured at 260 nm and 280nm using quartz crystal cuvette.

Quality Assessment

A ratio of OD values at 260nm and 280nm indicates the purity of the extracted DNA sample. If the ratio is within range of 1.6 to 2.0, then DNA sample is considered as clear and free from contaminants like residual protein and mRNA. An OD ratio less than 1.6 indicate the residual proteins or phenol contamination, whereas ratio of more than 2.0 indicates residual RNA contamination.

Quantity Assessment

DNA quantity was estimated as the OD value at 260nm of extracted sample is 1.00 then the concentration of the DNA is 50ug/ml.

Therefore, DNA concentration = OD at 260nm x 50 x Dilution factor.

DNA quantification and electrophoresis

Electrophoretic analysis of DNA using agarose gels can confirm DNA integrity. Typically intact genomic DNA will be up to 40KB in size, depending upon the species. Prepare 1% agarose gel has to be by adding required quantity of agarose to 1X Tris-Acetate-EDTA (TAE) buffer and mix well. Heat the mixture in microwave oven until it became clear and take care to avoid over boiling and evaporation. Cool the mixture to ~50^0 C and add ethidium bromide to make a final concentration of 0.001ug/ml. Pour the entire mixture into a tray in which combs are fixed to make wells in the gel. After gel formation, place the tray in buffer tank containing 1X TAE buffer for submerged gel electrophoresis and remove the combs with care to avoid rupture of wells. Mix 1µl of each DNA sample with 1µl of loading dye and load the mixture into the wells. Subject the Gel to electrophoresis at 90V for 30 minutes and visualise using gel documentation system where it is exposed to Ultraviolet rays. Under Ultraviolet rays exposure, DNA will give luminance which indicate the presence of DNA in the sample as shown in the below figure.

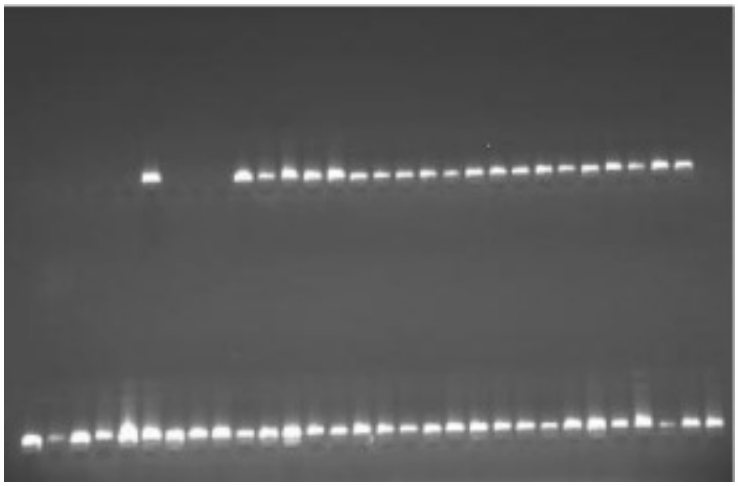


Fig. 2.6: Electrophoresis gel picture showing DNA bands

2.5.1 Polymerase Chain Reaction

PCR (Polymerase Chain Reaction) is a revolutionary method developed by Kary Mullis in 1980s. It is an essential and ubiquitous tool in genetics and molecular biology. With the use of this technique we can clone DNA *invitro*.

PCR is based on using the ability of DNA polymerase to synthesize new strand of DNA complementary to the offered template strand. Because DNA polymerase can add a nucleotide only onto a pre existing 3'-OH group, it needs a primer to which it can add the first nucleotide. This requirement makes it possible to delineate a specific region of template sequence that the researcher wants to amplify. At the end of the PCR reaction, the specific sequence will be accumulated in billions of copies (amplicons). DNA sequencing by dye termination technique requires multiple copies of DNA, hence PCR is performed to generate numerous copies of DNA fragments of interest which were further used for sequencing.

The genomic DNA of each subject can be amplified on thermal cycler with an initial denaturation at 96^0C for 3 minutes and later on for 35 cycles at 95^0C for 60 seconds, at estimated annealing temperature of the primer for 45 seconds, extension at 72^0 C for 2.30 minutes and a final extension at the end of 35^{th} cycle at 72^0C for 7 minutes in a final volume of 10 µl containing 50mM KCl, 10mm Tris (pH 8.3), 1.5mM MgCl2, 75 ng of each primer, 100 µM deoxy-NTP, and 1 U *Taq* polymerase.

Primer designing

The main limitation of PCR technique is to provide short pieces of single-stranded DNA (primers) that are complementary to a part of target sequence. With the use of human genome sequence available we can now design primers to any region of interest in the human genome. The most critical step in PCR experiment is designing oligo-nucleotide primers. As poor primers could result in little or even no PCR product, alternatively they could amplify unwanted DNA fragments. Either will affect the downstream analysis. Many of the factors which affect the primers specificity and sensitivity like product size, primer size, tm, GC content, GC clamps and dimer formation can be adjusted as per the user requirement. Primers which fit the specified criteria can be checked for their specificity using NCBI BLAST.

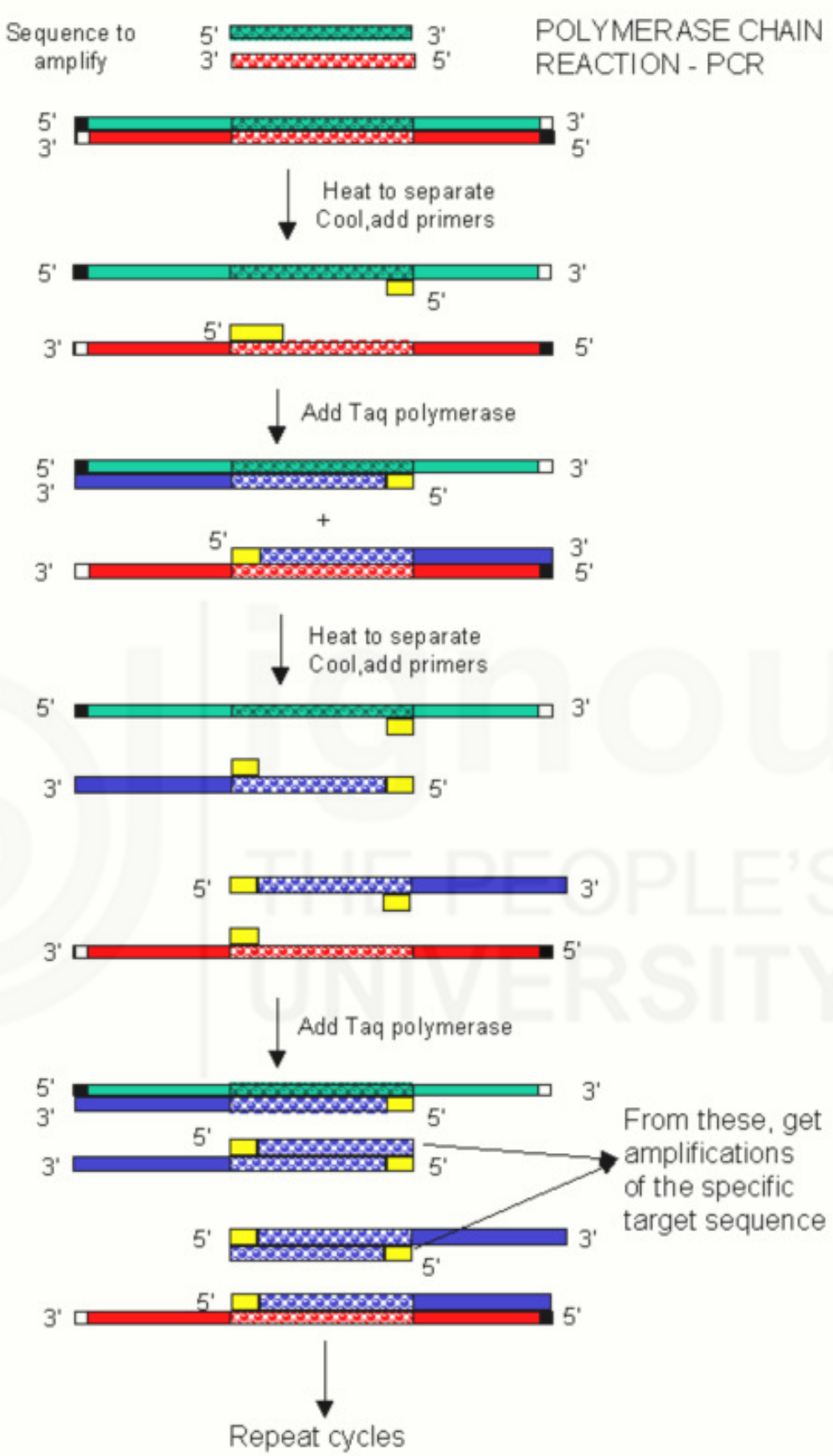


Fig.2.7: Schematic diagram of Polymerase chain reaction

2.5.2 Restriction Fragment Length Polymorphism

Restriction-Fragment Length Polymorphism (RFLP) was proposed by American geneticist David Botstein, biochemist Ronald W. Davis, population geneticist Mark Skolnick, and biologist Ray White. Restriction fragment length polymorphisms (RFLPs) can be used to produce a linkage map of the human genome and to map the genes that cause disease in humans.

Restriction Fragment Length Polymorphism (RFLP) analysis measures fragments of DNA containing short sequences that vary from person to person, called VNTRs. After extracting DNA from a sample and amplifying it with the technique known as Polymerase Chain Reaction, we can add restriction enzymes that cut the DNA at specific points. The resulting fragments can be sorted by length with gel electrophoresis technology to determine how many times a given VNTR is repeated.

If two different samples show VNTRs of different lengths, the samples could not have come from the same person. On the other hand, two samples showing VNTRs of the same length could have come from the same person, or from two people who happen to have VNTRs of the same length at that location. By comparing enough VNTRs from two individuals, however, the likelihood of a coincidental match can be reduced to nearly zero. RFLP testing requires hundreds of steps and weeks to complete, and it has been largely replaced by newer, faster techniques.

Fig. 2.8: Restricted enzyme EcoR-I identifies its specific GAATTC sequence and Cuts between G and A of the DNA strand. a) Showing the cutting positions. b) Showing resulted four strand fragments after enzyme digestion

Short Tandem Repeats in DNA Analysis

STRs can be amplified and sequenced using PCR and Sequencing techniques. Analysis is based on the number of repeats present in the sample.

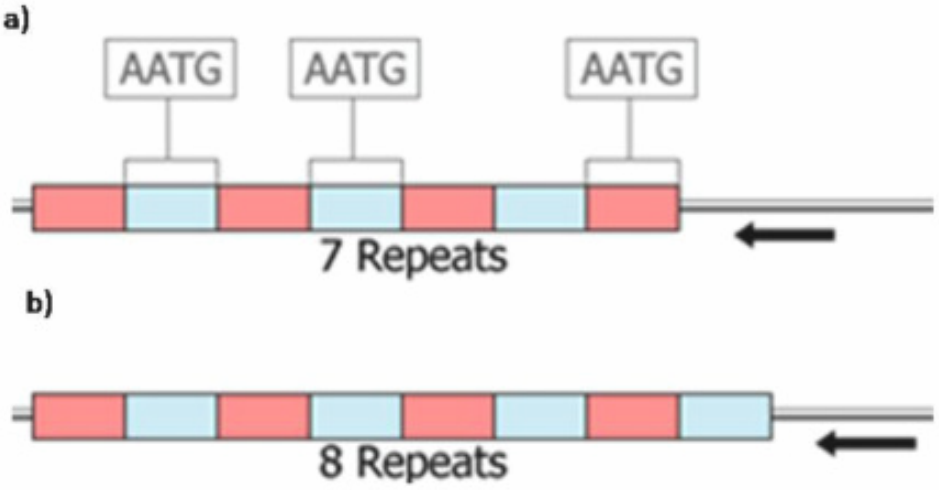


Fig.2.9: a) One DNA strand showing 7 repeats (STRs) of AATG. b) Another DNA strand showing 8 repeats of AATG.

2.5.3 DNA Sequencing Methods

In the 1960s and 1970s, British scientists Frederick Sanger and Alan Coulson, Alan Maxam and Walter Gilbert in the United States, developed DNA sequencing techniques. Automated equipment makes DNA sequencing a speedy, routine laboratory procedure. Sanger and Gilbert won the 1980 Nobel Prize in Chemistry for their work.

Sanger method of DNA Sequencing

In Sanger method, specific terminators of DNA chain elongation 2', 3'-dideoxynucleoside triphosphates are synthesized. These molecules can be incorporated normally into a growing DNA chain through their 5'- triphosphates groups. However, they cannot form phosphodiester bonds with the next incoming deoxynucleotide triphosphates (dNTPs). When a small amount of a specific dideoxy NTP is included along with the four deoxyNTPs normally required in the reaction mixture for DNA synthesis by DNA polymerase, the products are a series of chains that are specifically terminated at dideoxy residue. This forms the basis for Sanger's method.

Procedure

Initially single strand DNA is prepared through denaturation process. Then single strand DNA is mixed with a short end labeled piece of DNA (Primer) that is complementary to the end of single strand DNA. Labeling of primer is carried out using enzymes like Alkaline Phosphatase and Polynucleotide Kinase. After primer is annealed to DNA, sample is divided into four portions in four tubes. In each tube, along with DNA, Primer, DNA polymerase, a carefully controlled ratio of one particular dideoxynucleotide with its normal deoxynucleotide, and the other three dNTPs are added.

In each tube, DNA polymerase polymerizes normally from primer by utilizing nucleotides. When ddNTP is incorporated, the growth of that chain will stop. If the correct ratio of ddNTP: dNTP is chosen, a series of labeled strands will result, the lengths of which are dependent on the location of a particular base relative to the end of the DNA.

After suitable time period, the resultant labeled fragments in each tube are separated by size on an acrylamide gel. The separated fragments are detected by exposure of the gel to x-ray film through the process of autoradiography. From

the band developed in each lane of the autoradiograph and knowledge of which lane contain which base, the sequence of the complementary sequence can be obtained. From the complementary sequence, the sequence of the original strand can be easily determined with the help of Watson and Crick base pairing rule. Thus Sanger method is used for DNA sequencing.

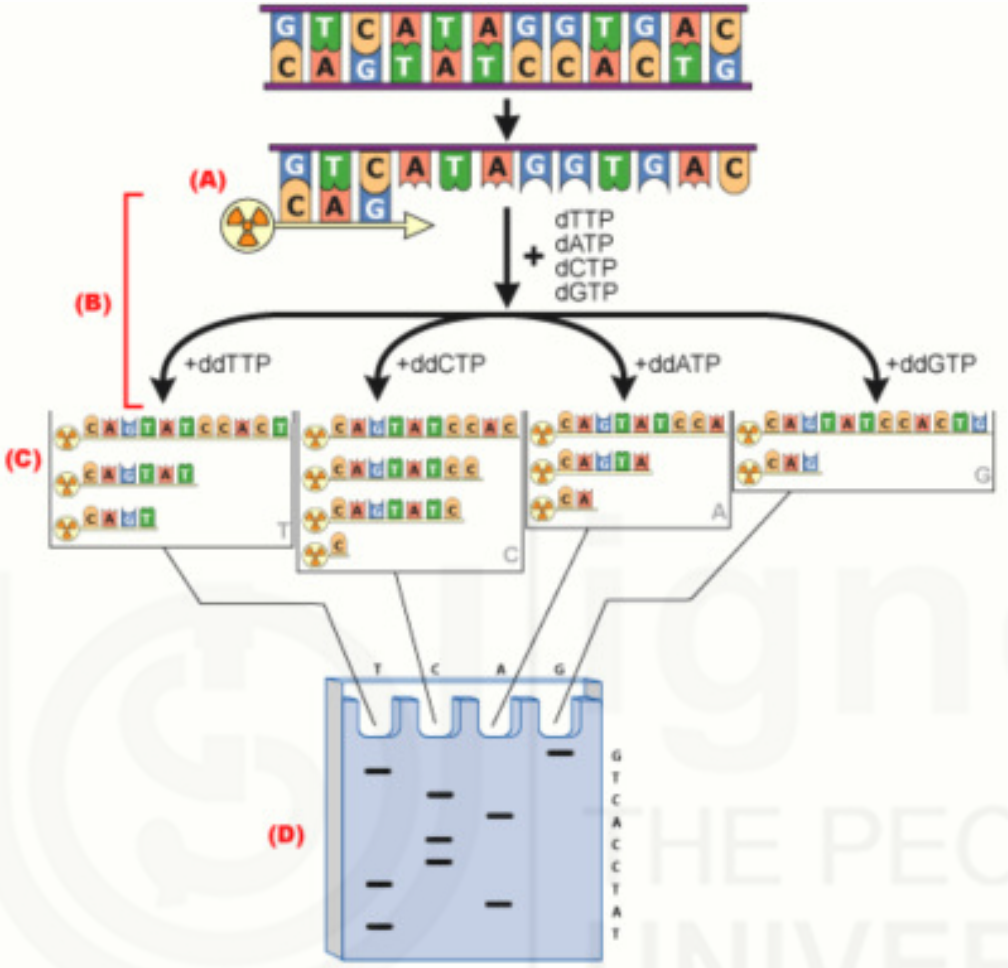


Fig.2.10: Schematic diagram of Sanger's enzymatic sequencing

Automated DNA Sequencing

There are various methods available for DNA sequencing like chemical degradation, chain termination method, sequencing by ligation and micro fluidic Sanger sequencing etc. Advances in automation have opened gates to new fast and reliable automated DNA sequencing technologies. Owing to its greater expediency and speed, dye-terminator sequencing is now the mainstay in automated sequencing. Dye-terminator sequencing is a slight modification of the Sanger's chain termination method it utilizes labelling of the chain terminator ddNTPs, which permits sequencing in a single reaction. In dye-terminator sequencing, each of the four dideoxynucleotide chain terminators is labelled with fluorescent dyes, with different wavelengths of fluorescence and emission. The dye labelled DNA fragments will be capillary electrophoresed and a detection system will identify the labelled bases when they pass through a laser that activates the dye.

Cycle sequencing

To read the sequence of the amplified DNA termination with florescent dye labelled ddNTPs, each base PCR amplicon should be subjected to cycle reaction with one primer using BigDye terminator cycle sequencing ready reaction kits following the manufacturer's guidelines.

Sequencing run

The PCR product should be added 10µl of Hi-Di formamide before feeding it to the sequencer machine.

Sequence Alignment

The generated sequences can aligned to their respective reference sequences with the use of software called DNA baser. It performs sequence comparisons for variant identifications, SNP discovery and validation. It allows analysis of the re-sequenced data, comparing the consensus sequences to a known reference sequence. The reference sequences for the gene studied are obtained from NCBI Gen bank data base.

Sequence Editing and Mutation Scoring

We can score the mutations from the aligned sequences by checking electropherograms of the DNA sequences. Genotypes can be exported from the software for further analysis.

Single Nucleotide Polymorphisms or SNPs (pronounced snips") are variations in a DNA sequence that occur when a single nucleotide in the sequence is different from the normal in at least one percent of the population. When SNPs occur inside a gene, they create different variants or alleles, of that gene.

Unlike repeated portions of DNA like STRs and VNTRs, in the case of SNPs it is the sequence itself, not its length that is useful to forensic scientists. SNPs are commonly occurring every 100 to 300 bases along the entire length of the human genome. Mutations in SNPs are very rare, so the sequences tend to be passed unchanged across generations. But because any given SNP is relatively common in the population, an analyst must examine dozens of SNPs to derive a true DNA fingerprint. For this reason, SNP analysis is rarely used in forensic cases.

Figure: Sequence showing Homozygous mutation

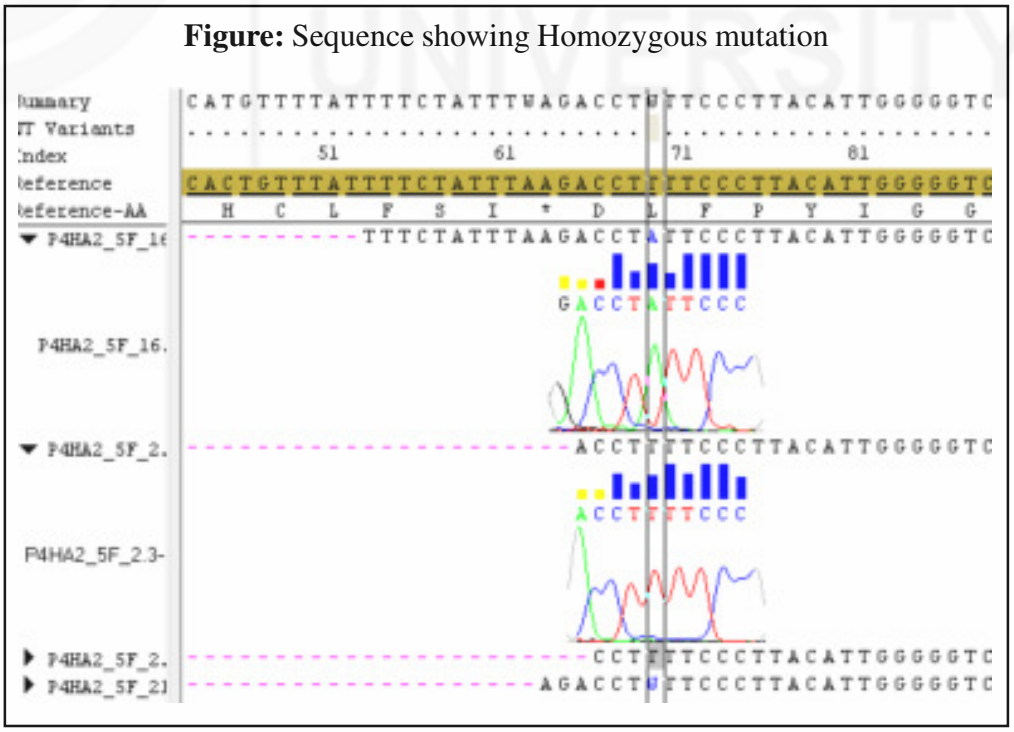


Fig.2.11: DNA Sequence Electropherograms showing mutations

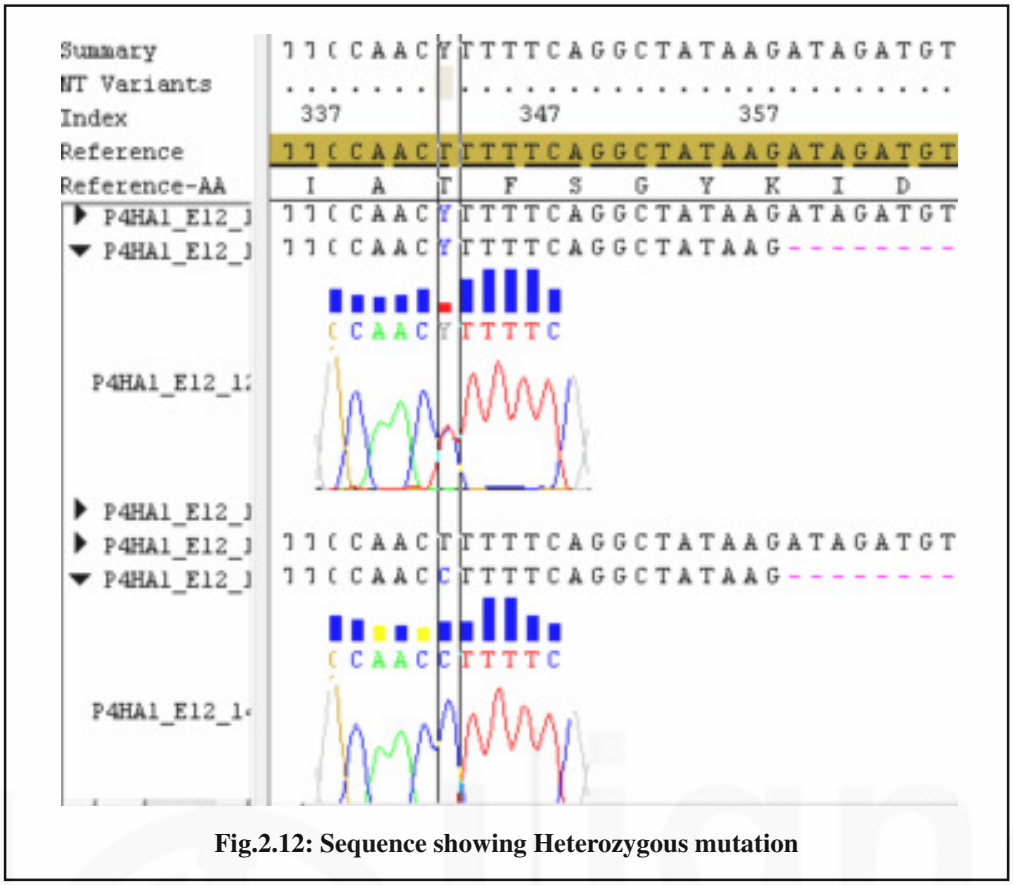


Fig.2.12: Sequence showing Heterozygous mutation

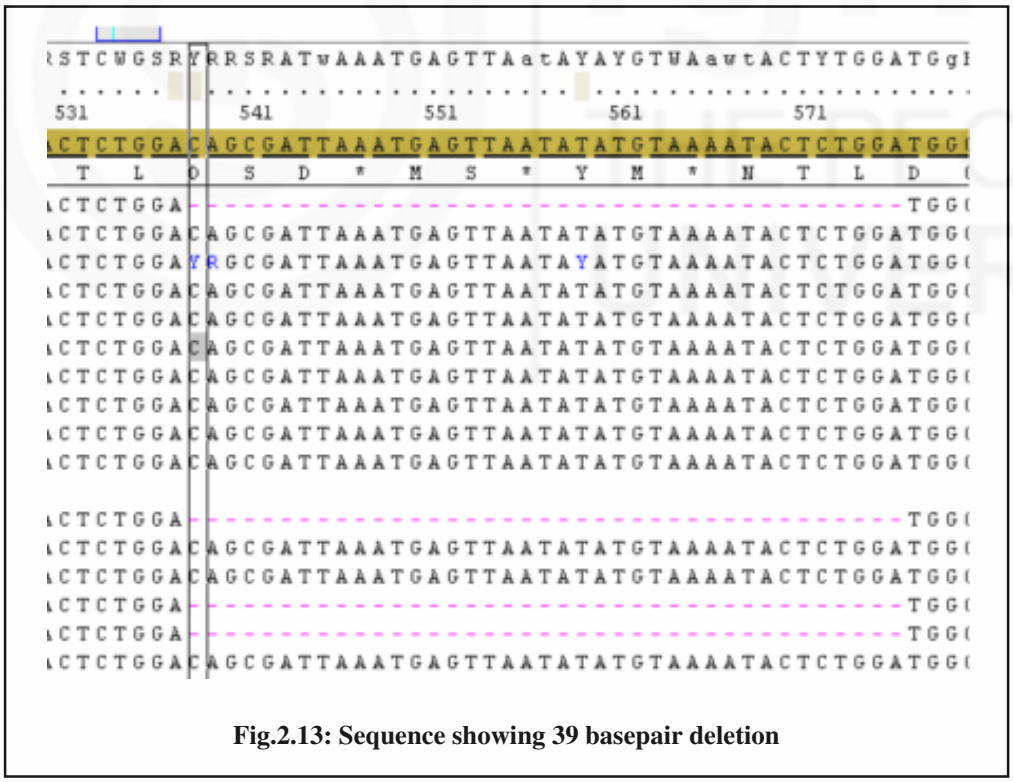


Fig.2.13: Sequence showing 39 basepair deletion

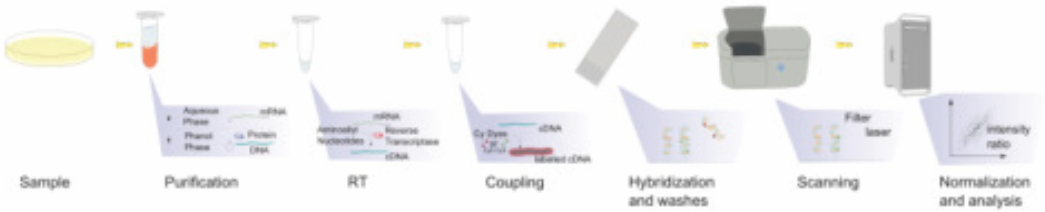
2.5.4 Microarray

A DNA microarray (also commonly known as gene chip, DNA chip, or biochip) is a collection of microscopic DNA spots attached to a solid surface. Scientists use DNA microarrays to measure the expression levels of large numbers of genes simultaneously or to genotype multiple regions of a genome. Each DNA spot contains picomoles (10-12 moles) of a specific DNA sequence, known

as probes (or reporters). These can be a short section of a gene or other DNA element that are used to hybridize a cDNA or cRNA sample (called target) under high-stringency conditions. Probe-target hybridisation is usually detected and quantified by detection of fluorophore, silver, or chemiluminescence-labelled targets to determine relative abundance of nucleic acid sequences in the target

DNA microarrays can be used to measure changes in expression levels, to detect single nucleotide polymorphisms (SNPs), or to genotype or resequence mutant genomes (see uses and types section).

Principle



Hybridisation of the target to the probe

The core principle behind microarrays is hybridisation between two DNA strands, the property of complementary nucleic acid sequences to specifically pair with each other by forming hydrogen bonds between complementary nucleotide base pairs. A high number of complementary base pairs in a nucleotide sequence means tighter non-covalent bonding between the two strands. After washing off of non-specific bonding sequences, only strongly paired strands will remain hybridized. So fluorescently labelled target sequences that bind to a probe sequence generate a signal that depends on the strength of the hybridisation determined by the number of paired bases, the hybridisation conditions (such as temperature), and washing after hybridisation. Total strength of the signal, from a spot (feature), depends upon the amount of target sample binding to the probes present on that spot. Microarrays use relative quantisation in which the intensity of a feature is compared to the intensity of the same feature under a different condition, and the identity of the feature is known by its position.

Many types of arrays exist and the broadest distinction is whether they are spatially arranged on a surface or on coded beads: (i) The traditional solid-phase array is a collection of orderly microscopic "spots", called features, each with a specific probe attached to a solid surface, such as glass, plastic or silicon biochip (commonly known as a genome chip, DNA chip or gene array). Thousands of them can be placed in known locations on a single DNA microarray. (ii)The alternative bead array is a collection of microscopic polystyrene beads, each with a specific probe and a ratio of two or more dyes, which do not interfere with the fluorescent dyes used on the target sequence.

DNA microarrays can be used to detect DNA (as in comparative genomic hybridisation), or detect RNA (most commonly as cDNA after reverse transcription) that may or may not be translated into proteins. The process of measuring gene expression via cDNA is called expression analysis or expression profiling.

2.6 SUMMARY

DNA, or deoxyribonucleic acid, contains genetic information in the form of genetic code that is responsible for physical appearance of the organism, biological development and maintenance of life processes. Any change in the DNA leads to different form of DNA in the population. Occurrence of more than one form of DNA in the population is called as DNA polymorphism. This Polymorphism may be due to variation at single nucleotide base, an array of nucleotide bases or at chromosomal level. DNA polymorphisms are playing an important role in understanding human evolution and in unravelling the genetic basis of diseases.

References

Rao V.R. and A.R.Sankyan, 2007. *Human origins, genome and people of India*. Anthropological Survey of India. Allied publications Pvt Ltd, New Delhi .

Ingman M, Kaessmann H, Pääbo S, Gyllensten U (2000) *Mitochondrial genome variation and the origin of modern humans*. Nature 408: 708–713.

Chandrasekar A, Satish Kumar, J. Sreenath, B. N. Sarkar, B.P. Urade, S. Mallick, S. S.Bandopadhyay, Pinuma Barua, S. S. Barik,Debasish Basu, U.Kiran, P. Gangopadhyay, Ramesh Sahani, B. V.Ravi Prasad, S.Gangopadhyay, G. R. Lakshmi, R. R. Reddy , K. Padmaja, P. N. Venugopal, M. B. Sharma, and V.R.Rao. *Updating Phylogeny of Mitochondrial DNA Macrohaplogroup M in India: Dispersal of Modern Human in South Asian Corridor.* PLoS ONE 4(10): e7447. doi:10.1371/ journal.pone.0007447(2009).

Barik, S.S. R. Sahani, B.V.R. Prasad, P. Endicott, M. Metspalu, B.N. Sarkar, S. Bhattacharya, P.C.H. Annapoorna, J. Sreenath, D. Sun, J.J. Sanchez, S.Y.W. Ho, A. Chandrasekar, V.R. Rao. *Detailed mtDNA genotypes permit a reassessment of the settlement and population structure of the Andaman Islands*. American Journal of Physical Anthropology. 136(1):19-27 (2008).

Chandrasekar, A et al. *YAP insertion signature in South Asia*. Ann Hum Biol. 34(5):582-586 (2007).

Palanichamy MG, Sun C, Agrawal S, Bandelt H-J, Kong Q-P, et al. 2004. *Phylogeny of mitochondrial DNA macrohaplogroup N in India, based on complete sequencing: implications for the peopling of South Asia.* Am J Hum Genet 75: 966–978.

Web resources:

http://ycc.biosci.arizona.edu/

http://yhrd.org

http://mitomap.org

http://phylotree.org

http://projects.tcag.ca/variation/decipher

http://ncbi.nlm.nih.gov/projects/SNP/

Suggested Reading

Tom Strachan and Andrew P. Read. 2004. *Human Molecular Genetics*. Garland publishing, Taylor and Francis Group, London and New York.

Sambrook, J. and Russell, D. 2001. *Molecular cloning: A laboratory manual*, 3rd Edn, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York.

Human Genome Nature Issue (15 February 2001) *Nature* 409, 813-958. http://www.nature.com/genomics/human/

Human Genome Science Issue (16 February 2001) *Science* 291, 1177-1351. http://www.sciencemag.org/content/vol291/issue5507/index.shtml

U.S. National Human Genome Research Institute: www.ncbi.nlm.nih.gov/

About human genome project: http://www.genome.gov/10001772

Human genome Browsers and Integrated Databases: www.ensembl.org/

Sample Questions

1) Define what in DNA polymorphism and its forms?
2) Explain how mtDNA polymorphism helped in understanding modern humans in Indian subcontinent?
3) Discuss the Y-chromosome tree?

Short Notes

1) Polymerase Chain Reaction?

   Sanger method of DNA sequencing?

   RFLP and its uses?

   Microarray and its usage?

UNIT 3 HUMAN GENOME PROJECT

Contents

Learning Objectives

After studying this unit, you would be able to:

- explain and appreciate the significance and outcome of the biology's largest programme – The Human Genome Project;
- to discover the secrets of life in terms of genetic makeup of biological systems;

- to understand the resemblance and differences between the humans and other organisms in terms of sequence variations;
- to explain the genetic differences between world populations and evolution of mankind; and
- discuss the applications of sequence information for the benefit of mankind and society in general.

3.1 INTRODUCTION

The life processes in any living organism are controlled by a set of genes that are located on chromosomes which are present in numbers that are highly specific for a given species. In humans there are 23 chromosomes present as pairs in all somatic cells (referred as diploid number or 2n) and as a single unit in gametic cells (referred as haploid number or n). Of the 23 pairs of chromosomes present in an individual, one set is inherited from the father and another from the mother along with the genes carried by them. Hence we see the resemblance of characters between the parents and their children. In human system there are trillion cells of different types that are organised into various tissues/organs that carry out myriad functions related to day to day life processes. All the functions carried out by these cells are controlled by genes located on the 23 chromosomes (table-3.1a and 3.2b).

Table 3.1: Number of Entries in Online Mendelian Inheritance in man (OMIM) as reported on December 9, 2011

	Autosomal	X-Linked	Y-Linked	Mito-chondrial	Total
* Gene with known sequence	13017	638	48	35	13738
+ Gene with known sequence and phenotype	171	6	0	2	179
# Phenotype description, molecular basis known	3048	258	4	28	3338
% Mendelian phenotype or locus, molecular basis unknown	1654	136	5	0	1795
Other, mainly phenotypes with suspected mendelian basis	1800	129	2	0	1931
Total	**19690**	**1167**	**59**	**65**	**20981**

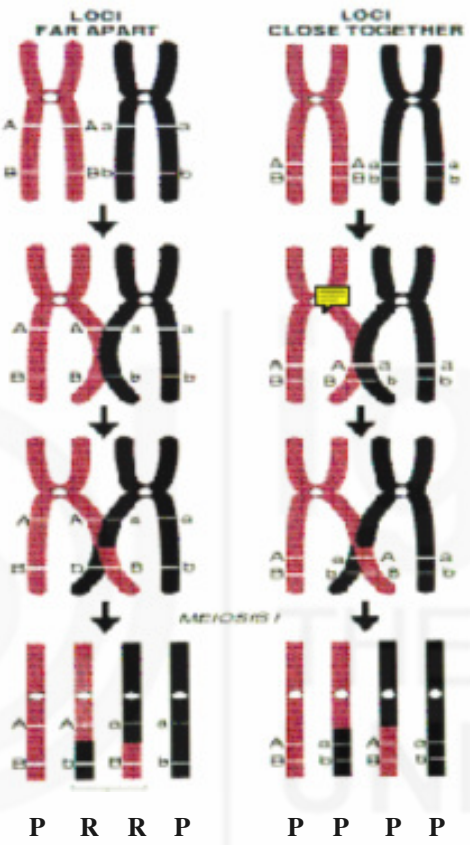
Table 3.1b: Synopsis of human gene map

Chromosome	Loci	Chromosome	Loci	Chromosome	Loci
1	1288	9	494	17	756
2	837	10	483	18	187
3	695	11	809	19	830
4	503	12	685	20	333
5	611	13	246	21	143
6	783	14	419	22	327
7	598	15	392	X	716
8	466	16	529	Y	46
				Total	**13176**

Physical location of all the genes on different chromosomes of an organism is represented as "Genetic Maps". These maps are generated by determining distances between different genes present on a chromosome by an approach called linkage analysis [Box-3.1]. The distance between two genes is expressed as CentiMorgan (cM) which is a unit of genetic distance. Thus 1cM represents 1% probability of recombination occurring during meiosis i.e. gametogenesis. The genes present on the same chromosome are said to be linked and the group of such linked genes are called linkage groups or chromosomes.

Genetic Mapping

a) Recombination



P: Parental combination of genes; R: Recombinants resulting due to crossing over occurring between the loci A and B

b) Linkage estimation

Maximum likelihood estimate (mle) of recombination fraction (referred as è or x or r) is used for determining linkage between two loci on a chromosome based on the relative probability (P_R) of having obtained the family. The P_R is determined by calculating the probability of having obtained the various combinations of the particular trait under consideration on the assumption of there being no measurable linkage (H_0= No linkage; è = 0.5) and comparing this with the probabilities based on a range of recombination fractions è varying from 0.00 to 0.05, i.e.

$$P_R = \frac{\text{P (FAMILY, GIVEN è = 0.0 to 0.5)}}{\text{P (FAMILY, GIVEN è = 0.0 to 0.5)}}$$

P_R is expressed as its logarithm. The $\log_{10}$ of PR is called as log of the odds or lod score. Lod score value of = 3.0 indicates that the two loci tested are linked, value of 2.0 as evidence of strong linkage, value of 1.0 as evidence for tentative linkage and that of -2.0 absence of linkage. The Lod score of 2.0 that is suggestive of linkage can be further evaluated by analyzing more candidate/marker loci and by screening additional members of the family.

Chemically genes are made up of a macromolecule called Deoxyribose Nucleic Acid (DNA) which exists as a double helical structure resembling a ladder. Chemically, DNA comprises 4 nitrogenous bases- Adenine (A), Guanine (G), Thymine (T) and Cytosine (C) which are arranged as rungs of the ladder and supported by a sugar-phosphate backbone (Fig.3.1). Each base with a sugar and phosphate molecule is referred as a nucleotide. The nitrogenous bases Adenine and Guanine are referred as "Purines" and Thymine and Cytosine as "Pyrimidines". This structure of DNA as described by Watson and Crick (1953) satisfies all the criteria of a genetic material including the segregation of different genes/characters through generations.

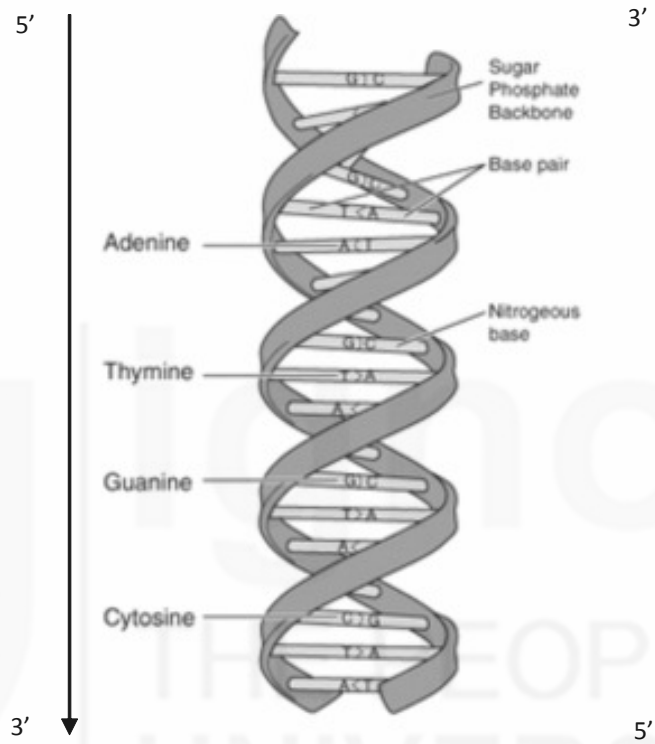


Fig.3.1: Double helical structure of the genetic material

The structure of de-oxyribose nucleic acid (DNA) comprising 4 nitrogenous bases (Adenine, Guanine, Thymine, Cytosine) each attached to a sugar and phosphate molecule that form a backbone. Adenine and Guanine are called purines and Thymine and Cytosine as pyrimidines. The base pairing is strictly complementary i.e. Adenine always pairs with Thymine while Guanine always pairs with Cytosine. The bases are held together by hydrogen bonds forming rungs of the ladder. The two strands of the DNA forming twisted double helical structure run in the opposite directions i.e. one strand runs from 5' to 3' and the other from 3' to 5'direction.

In early 1950s human geneticists have attempted to map some disease genes using certain genetic markers like ABO blood groups. One such study established close linkage between the loci for a disease called Nail-Patella syndrome and that of ABO blood groups. In the following years researchers attempted to map several other disease genes using different polymorphic loci [Box-3.2] related to serum proteins, enzymes and leucocyte (HLA) antigens. Later with the discovery of enzymes called restriction endonucleases or restriction enzymes (REs), new markers known as restriction fragment length polymorphisms (RFLPs) were identified which proved to be better markers in genetic analysis experiments.

Restriction enzymes cut the DNA at specific sites breaking them into fragments of different sizes. If a given DNA sequence has one restriction site 2 fragments of different lengths will be generated. The number of fragments generated will be n+1 when n number of restriction enzymes is used to cut the given DNA sequence. In later years certain sequences of nucleotides or base pairs (bps) were found to be repeated in various numbers differing in different individuals thus showing polymorphism. These stretches of repeats (0.1-20kb long) are referred to as variable number of tandem repeats (VNTRs) or minisatellites where the core repeat sequence of DNA carries 15 to hundreds of nucleotides. Initially DNA finger printing – a technique followed in forensic science used certain VNTR markers. Later smaller stretches of repeat sequences (<0.1 kb) called microsatellites with only 1-4 nucleotides (occurring as di, tri and tetranucleotide repeats) in each stretch were identified which are highly polymorphic. Now more than 6000 such markers located on different chromosomes are available for conducting any study. VNTRs and microstaellite markers were extensively used in 1990s in gene mapping studies and studies on their associations with diseases and risk predictions. With the discovery of single nucleotide polymorphisms (SNPs read as snips) which distinguishes individuals at single nucleotide level, research in human genetics took a different turn with the application of genome wide screening for mapping genes, finding differences between population groups, between normal and disease samples which in turn aids in formulating better treatment measures.

Genetic Polymorphisms

Genetic polymorphisms is defined as the presence of more than two allelic forms at a given locus found in individuals in such frequencies that the rarest of them does not occur just due to recurring mutations but it is due to a phenomenon called "polymorphisms". The frequency of the rarest allele/form as a rule is taken as > 1.0%. Several genetic loci related to red cell antigens, serum proteins, enzymes, leucocyte antigens and DNA markers have been identified over the years which are used in gene mapping, studies on associations with diseases and risk prediction and also tracing the origin of human population groups and estimating genetic distances. The different polymorphic markers available for such studies are given below

Type of Marker	No. of Loci	Features
Blood Group antigens	>25	Genotype cannot always be inferred from the phenotypes. Difficult for physical localization of genes
Serum Proteins	30	Often limited polymorphisms. Difficult for physical localization of genes
Leaucocyte antigens (HLA)	1	HLA system with A,B,C,D and DR loci each harbors hundreds of alleles thus resulting extensive polymorphism. ls highly informative useful in gene mapping
Restriction Fragment Length Pol-ymorphisms (RFLPs)	$>10^5$	Bi-allelic markers, maximum heterozygosity 0.5, genotyped using Southern blotting and PCR techniques. Easy for physical localisation
Variable number of tandemn rep-eats (VNTRs)	$>10^4$	Many alleles at each locus. Highly informative. Easy for physical localization of genes. Tend to cluster near the ends of chromosomes
Microsatellites (di, tri and tetra-nucleotide repeats)	$>10^5$	Many alleles at each locus. Highly informative. Useful for physical localization of genes. More than 6000 markers identified that are distributed throughout the genome. Sometimes unstable.
Single Nucleotide Polymorphisms (SNPs)	$>4x10^5$	Shows variation between any two individuals at single nucleotide level. Are most stable but less informative than microsatellites. Can be genotyped on a very large scale using automated sequencing. Useful for genome wide screening, gene mapping and risk prediction for diseases

In spite of the availability of thousands of polymorphic markers (RFLPs, VNTRs, Microsatellites and snips), generation of human genetic map by mapping all the estimated 30,000 genes one by one appeared to be a Herculean task in that it is both tedious and time consuming. Hence to overcome this difficulty, the idea of determining the entire sequence of nucleotides in the DNA was floated which was discussed in depth at several scientific meetings before it was finally approved and the stage to undertake the Human Genome Project (HGP) was set.

What is a genome?

A genome represents total set of different DNA molecules (DNA content) including all of its genes along with spaces between them in an organelle (like mitochondria), cell or an organism. The genome in a species is organised in a specific manner with features to co-ordinate various functions and also reproduction to keep up with continuity of the species. Each genome is a blue print that contains all of the information needed to build and maintain an organism. Human genome is more complex with variation in its organisation found in the nucleus and mitochondrial components. A complete sequence of human mitochondrial (mt) genome was published in 1981 By Fred Sanger and his colleagues.

3.2 HUMAN GENOME PROJECT (HGP)

The human genome project was an international effort to sequence every nucleotide in the human genome and to identify all the genes contained within the genome. This effort was coordinated by United States Department of Energy and National Institutes of Health (NIH). It was the highest ever funded programme in biology and laboratories from UK, Japan and Germany were also associated with it.

3.2.1 History of HGP

The implementation of HGP was not instantaneous, but it was the outcome of careful efforts put forward by the scientists after several deliberations. In a way the idea of human genome sequencing was initiated in 1977, when the dideoxy DNA sequencing method (Fig 3.2a and 3.2b) that was simple and efficient was discovered by Sanger and his colleagues from Cambridge, UK. In 1980 it became apparent that better understanding of biology of organisms will be achieved when the detailed structure of DNA base-by-base is understood. In 1984 for the first time US Department of energy (DOE) held a workshop at Alta, Utah to address the problem of detecting low frequency of very rare mutations in humans exposed to radiations and other environmental hazards. The meeting focused on the methods and technologies needed for the detection and characterisation of the mutations (sudden heritable changes occurring in a gene/DNA) for which, it was felt that the entire genome sequencing required. In the following years, the meetings held led to the formal proposal of need to sequence human genome to derive benefits in furthering cancer research (Mc Conkey, 1993).

A dedicated Human Genome Project was conceived in 1986 by DOE at Mexico Conference and objectives, cost involved, time needed etc. were discussed. In 1987 DOE's repot on human genome initiative has fore seen three major objectives 1) Generation of refined physical maps of human chromosomes 2) Development of support technologies and fecilities for human genetic research

and 3) expansion of network and increasing the computational and database capabilities. DOE responded to Santa Fe Meeting's report in 1987 and in 1988 National Institutes of Health (NIH) set up an office for Human genome research to co-ordinate genome research activities of NIH and other organisations. US congress authorized NIH and DOE to allocate funds for HGP. 3-5% of the budget was allocated for the programmes on ethical, legal and social implications (ELSI). In the same year an International Human Genome Organisation (HUGO) was founded by the efforts of independent group of scientists to coordinate national efforts, facilitate exchange of research resources, public debate and availability of information on the implications of human genome research and also sequencing and mapping of cDNA. While not involved in any research by itself, HUGO organises workshops for the benefit of researchers in the field.

Finally on 1st October, 1990 a 15 years programme with the budget of $3-billion to sequence Human Genome was officially launched by US DOE and NIH. In addition to US, universities and research centers from United Kingdom, France, Germany and Japan were involved in the project. As the commencement of the project picked up, altogether 18 different countries and companies participated in the programme. In the following year The Genome Data Base (GDB) repository was established for human DNA mapping data and it was made available for all working in the field. James D. Watson, co-investigator of DNA double helical structure was recruited as director of National Institute for Human Genome research (NIHGR) who visualised the creation of complete catalogue of three billion base pair in the human genome and mapping of all the genes. He continued to direct the project till 1992 and was replaced by Francis Collins who took over the charge of HGP project in 1993 which was renamed as National Human Genome Research Institute (NHGRI) in 1997. A parallel initiative was undertaken by Celera Corporation in 1998 and Dr. J. Craig Venter working for the corporation led the programme of sequencing Human Genome. To catch up with public funded government programme the company started a faster and cheaper approach with the cost of $300 million as against government funding of $3.0 billion.

The public and private ventures competed neck to neck and announced the first draft of human genome sequence simultaneously on June 26th, 2000. The first rough working draft was published in Nature by the public funded government project (2001) and in Science by Celera (Venter et. al., 2001). Before that in 1999 an international consortium of HGP comprising geneticists from UK, France, Germany, Japan, China and India announced the first complete sequence of human chromosome 21 (Hattori et. al. 2000). The draft covered about 83% of the genome, and with 90% of the euchromatin regions with 150,000 gaps and order and orientation of many segments still to be established. Finally ~ 92% of the human genome sequence was completed in 2003 two years earlier than the target date set at 2005. This was possible mainly because of the development of advanced and efficient technologies for automated sequencing and networking programmes that were supported by the project.

With additional funding, HGP also focused to sequence several other nonhuman organisms including the bacteria Escherichia*. coli.*, nematode, *Caenorhabditis elegans* the fruit fly, *Drosophila melanogaster*, the yeast *Saccharomyces cerevisiae*, the mouse *Mus musculus.* etc. Information on the sequences of different organisms facilitates comparative mapping studies and understanding the differences and similarities (homology) with human genome and sequences that are conserved among the species during evolution. Originally HGP developed haploid reference genome that comprises 3.2 billion nucleotides.

Any scientific problem starts with a hypothesis and objective followed by stepwise protocol finding the cause including sequencing of gene(s) concerned. But in case of HGP, it reverses the way in which any scientific project is conducted. It first aims at sequencing and then interprets the results later. In other words it first identifies the putative gene(s) based on the nucleotide sequence but will not identify their functions. Thus human genome studies do not end with sequencing the putative gene(s). It has to go through the tedious and challenging process of identifying the boundaries between the genes and other features from the raw DNA sequence which is called "Genome Annotation". The future lies in knowing the functions of the genes, assessing the interaction between genes and the environment and also correlating the observations made with developmental, biochemical and physiological processes going on in an organisms.

3.2.2 Goals of Human Genome Project

The goals set by Human Genome Project were:

- Identifying all of the estimated 30,000 genes in human DNA and mapping each gene to a site on one of the 23 chromosomes.
- Production of a variety of physical maps of all human chromosomes and that of selected organisms.
- Determination of the complete sequence of human nuclear genome and that of selected model organisms.
- Development of the capabilities for collecting, storing, distributing and analyzing the data generated.
- Creation of necessary technologies to meet the goals of the project.

In 1998 the following new 5 year goals were set

- Identification of the human genome variation between persons (i.e. single nucleotide variations between any two persons) since such variations are expected to play an important role in individual's response to infections, drugs and toxins.
- Comparison of human genomes with that of model organisms like bacteria, mouse, yeast, nematode, fruit fly, etc.
- Developing advanced computational capability to collect, store and analyse sequencing data.
- Addressing the ethical, legal and social implications (ELSI) concerned to the use of genetic tools and data.
- Developing interdisciplinary training programmes for future genomics researchers.

3.2.3 Strategies of Sequencing

The basis for the human genome sequencing was the dideoxy sequencing method developed by Fred Sanger and his colleagues in 1977. The basic principle of the technique remained the same in HGP programme but with improvements made regarding the efficiency by using the fluorescence labeled automated sequencers and capillary sequencers which helped in obtaining much higher sequencing throughputs (Fig. 3.2a and 3.2b). Dedicated computer programmes like PHRED, PHRAP were developed simultaneously which helped sequence interpretation, scanning for overlapping regions and data assembly.

Two different approaches were used to determine the first draft of genome sequence 1) Public funded project planned by International Human Genome

Sequencing Consortium (IHGSC) headed by Francis Collins and 2) Planned by the Celera Genomics led by Craig J. Ventor (Fig.3.3).

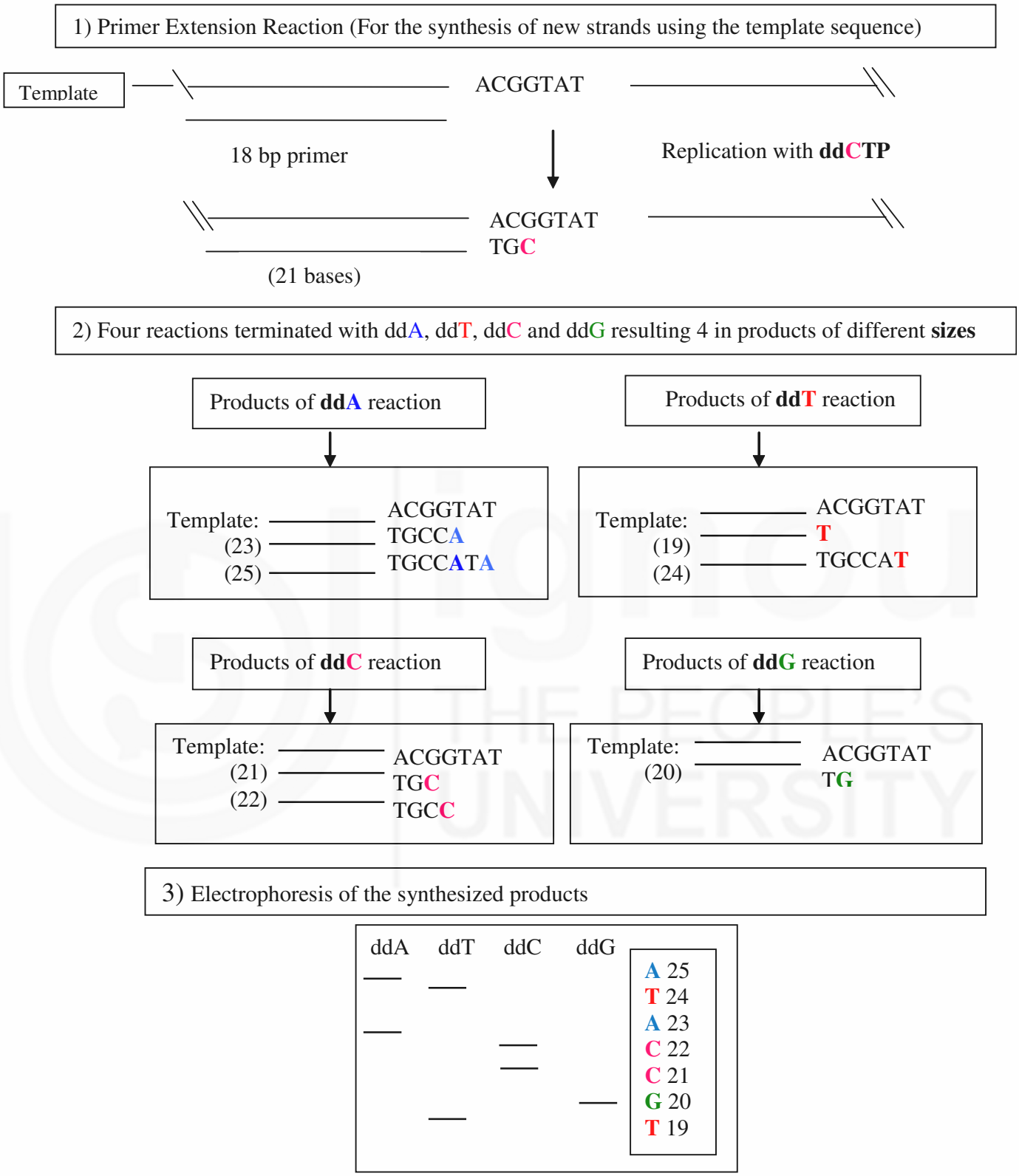


Fig.3.2a: Sanger's Method of DNA sequencing

Using the target sequence ACGGTAT, primer with 18 bps, radiolabelled dideoxy nucleotides (ddA, ddT, ddC and ddG) and polymerase enzyme new strands are synthesised. In the presence of dideoxy nucleotides the synthesis of new strands of DNA are terminated whenever the specific ddNTP is added. Thus products of different sizes are generated (that are of 19 to 25 bps length) which can be separated by gel electrophoresis. The original sequence (TGCCGT) can be read from the order/ladder of electrophoresis bands developed on the gel.

3.2.3.1 Approach of International Consortium

The public funded HGP was based on the "hierarchical shotgun" sequencing which involves random cleaving by sonification of starting DNA (from a chromosome) into several hundreds of fragments (150,000 bps in length) followed by end repair. These fragments are then cloned into a vector known as "Bacterial Artificial Chromosomes" or BACs which are derived from genetically engineered bacterial chromosomes. These vectors containing the genes or DNA fragments can be inserted into bacteria where they multiply using the bacterial DNA replication machinery. The BAC contents were known to correspond to specific locations on the chromosomes called sequence tagged sites (STSs). Several copies of each BAC were cut or 'shotgunned", into approximately 80 overlapping pieces which were then sequenced. A powerful computer programme was used to assemble the overlapping pieces into overall sequence for each chromosome. This process is nothing but mapping. The entire procedure is referred as "hierarchical shotgun" since the genome is first broken into relatively large pieces which are then mapped to chromosomes before being selected for sequencing.

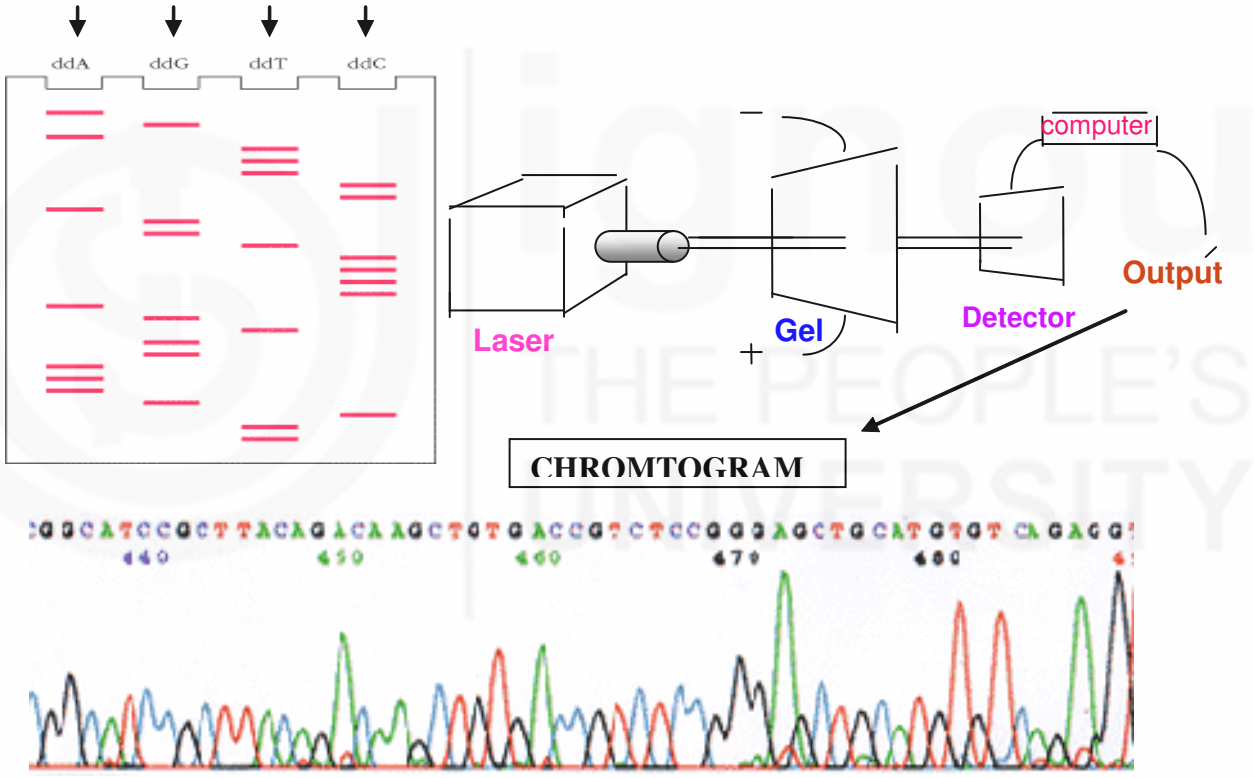


Fig .3.2b: Automated DNA Sequencing

The primer extension reactions carried out in automated sequencing is similar to that of Sanger's method except that the primers in each reaction are labeled with a different fluorescent staining molecule that emits light of a distinct color i.e. red for thymine, green for adenine, blue for cytosine and black for guanine. The different primer extension reaction products separate according to size upon gel electrophoresis. The bands are color coded. A laser beam that passes through the gel excites the fluorescent tag on each band and the detector analyses the color of the resulting emitted light. This information is converted into a sequence of bases and is stored in a computer. Print outs can be taken from the computer and the chromatogram will give the sequence details as peaks of different colors corresponding to the color of the fluorescence dye used for each base. In the

above diagram the sequence of nucleotides in 440- 446 positions are TCCGCTT that can be read by the color of the peaks.

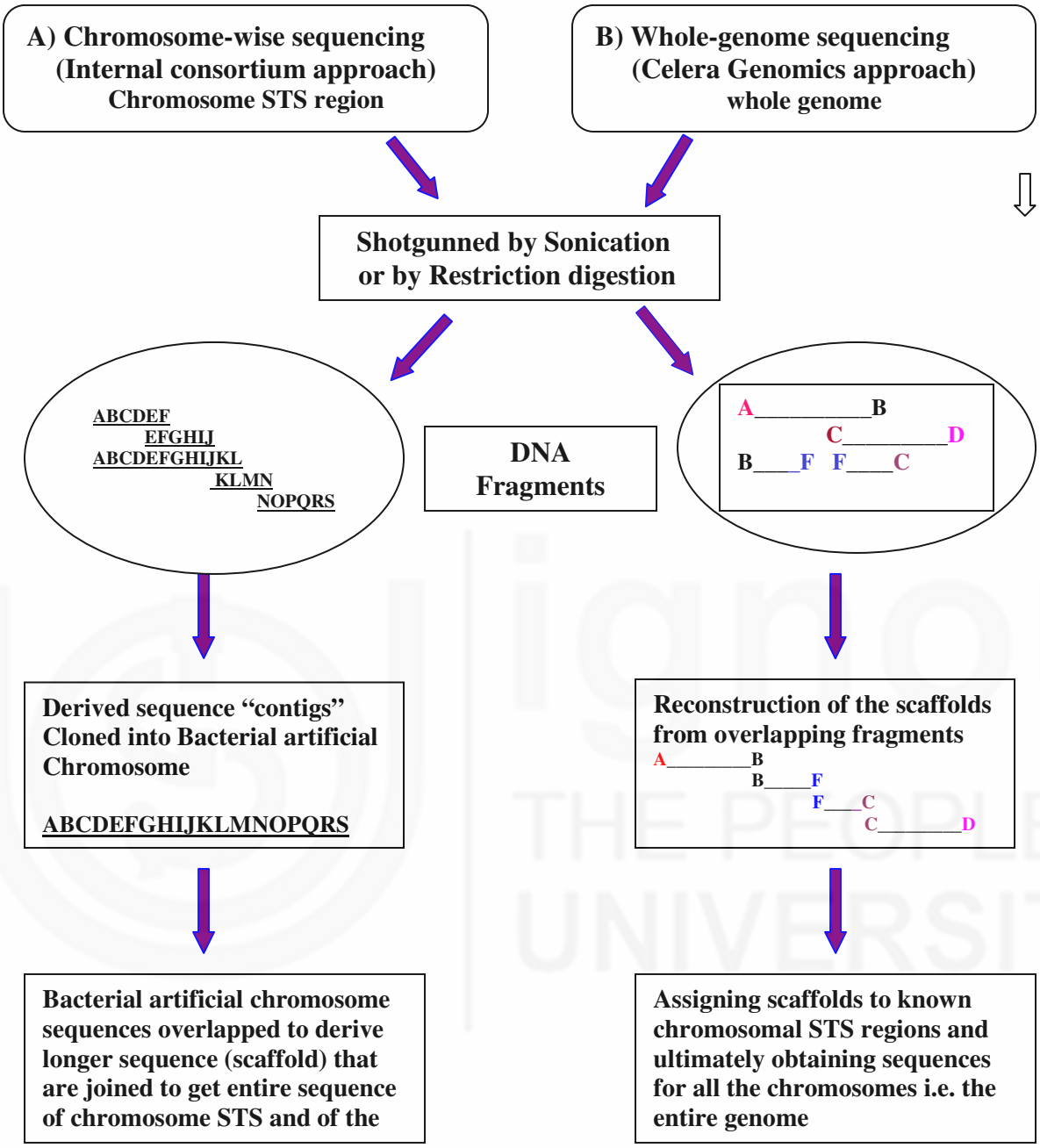


Fig. 3.3 : Startegies of human genome sequencing. Methods adopted by A) International human genome consortium and B) Celera Genomics. Instead of nucleotide symbols A,G,T and C alphabets A to S are used for convenience.

3.2.3.2 Approach of Celera Genomics

Celera Genomics headed by J. Venter followed "whole-genome shotgun" technique to sequence the human genome employing pairwise end sequencing. This technique was used to sequence bacterial genome of up to 6 million base pair in length, but not for large genome of 3.2 billion base pairs found in human genome. The technique skipped the BAC stage and used shotgunning multiple copies of the genome into small pieces. These pieces were then assembled into large overlapping sequences called "scaffolds" (frame work) using powerful computer programmes. There were 119,000 scaffolds which were assigned to

chromosomal sequence tagged sites (STSs). Celera company used information from public database but denied the access to any one to the private database generated by it. Celera's approach was rapid and of low cost involving only $ 3 millions as compared to publicly funded project of $3.2 billions.

3.2.4 Genome Donors for Sequencing

In the IHGSC, an international public-sector HGP, researchers collected samples of blood from females and that of sperm from males from large number of donors. Only a few of many of these (2 male and 2 female samples out of 20 each) were processed for DNA sequencing. Neither the donors nor the scientists knew the source of the samples and thus identity of the donors were protected. Much of the sequence (>70%) of the reference genome produced by the public HGP came from a single anonymous male donor from Buffalo, New York. For the Celera Genomics private-sector project samples were collected from 21 different individuals and only DNA of 5 individuals were used for sequencing.

3.2.5 Genome Assembly

Genome assembly which is a difficult computation method, is the process of arranging a large number of short sequences of DNA together to create a representation of the original chromosomes from which the DNA originated. In a shotgun sequencing project, all the DNA from an organism is first broken into millions of small pieces. These pieces are then "read" by automated sequencing machines, which can read up to 1000 nucleotides (with the bases adenine, guanine, thymine and cytosine). A genome assembly algorithm picks up all the pieces of DNA and aligns them to one another by detecting all regions where two of the short sequences, or "reads" overlap. These overlapping reads can be merged together, and the process continues. The draft genome sequence is produced by combining the sequenced contigs (ordered arrangement of cloned overlapping fragments) information and using linking information to create "scaffolds" (frame work). Scaffolds are then positioned or assigned to known chromosomal sequences tagged sites (STSs) creating a path.

3.2.6 Genome Annotation

Once the draft sequence is ready, Genome annotation has to be followed. Genome annotation is the process of attaching biological information to the sequences obtained. It is a major challenge for the HGP and covers a) structural annotation that deals with identification of genomic elements like open reading frames (ORFs), gene structure, coding regions and location of regulatory motifs and b) functional annotation that deals with attaching information about biological function, biochemical function, gene regulation and interactions and gene expression to the genomic elements. These steps involve both biological experiments and in silico analysis (bioinformatics).

Automatic annotation tools perform all the annotation by computer analysis.The basic level of annotation is using basic local alignment search tool (BLAST) for finding similarities between the sequences studied and then annotating genomes based on that. Genome annotation is an active area of investigation undetaken by different organisations which publish the results of their efforts in publicly available biological databases accessible via the web and other electronic means.

The HGP catalogued the information on the sequence of nucleotides in thousands of DNA fragments in a public database called GenBank maintained by US National Center For Biotechnology Information (NCBI) and sister organisations in Europe and Japan. From GenBank data base, sequences of known and hypothetical genes and proteins can be retrieved. The data bases are open for every one through internet. Other organisations like, Genome Bioinformatics group from University of California, Santacruz and Ensemble provide additional data and annotation and powerful tools for visualising and searching it.

3.2.7 Observations Drawn from Human Genome Sequencing

The draft genome sequence published in 2001 and the complete genome sequence published in 2004 reported the following findings:

- The human genome contains 3.1647 billion chemical nucleotide bases (A, C, T, and G).
- The total number of genes estimated at 30,000.
- The average gene consists of 3000 bases, but sizes vary greatly, with the largest known human gene being dystrophin at 2.4 million base.
- Almost all (99.9%) nucleotide bases are exactly the same in all people and the functions are unknown for over 50% of discovered genes.
- The number of genes in human beings is of same range found as in mice and round worms. Understanding how these genes express themselves and function would help to know how human diseases are caused.
- About 1.1% to 1.4% of the genome's sequence codes for proteins that carry out required functions in an organism.
- 98% of the genome is non-coding for proteins and misnamed as "junk DNA". Now much of the junk DNA is found to code for RNA which regulates other genetic and cellular functions.
- The human genome has high level of segmental duplications i.e. nearly identical, repeated sections of DNA than the other mammalian genomes. These repeated sequences may underline the creation of primate specific genes.
- Repetitive sequences are thought to have no direct functions, but they shed light on chromosome structure and dynamics.

Over time, the repeats reshape the genome by rearranging it, creating entirely new genes, and modifying and reshuffling existing genes. During the past 50 million years, a dramatic decrease seems to have occurred in the rate of accumulation of repeats in the human genome.

3.2.8 How is the Human Genome Arranged or Organised?

Genome sequencing facilitated better understanding of a) Nature of the genes controlling several traits b) Nature of Mutations resulting in altered functions of proteins c) Manipulation of the genome and predicting the consequences.

The human genome has gene-dense "urban centers" that are predominantly composed of the DNA building blocks G and C. In contrast, the gene-poor

"deserts" are composed richly of DNA building blocks A and T. Under the microscope GC- and AT-rich regions can be observed as light and dark bands on chromosomes representing euchromatin and heterochromatin regions. Genes appear to be concentrated in random areas along the genome, with vast expanses of long stretches of non-coding DNA between them. Stretches of up to 30,000 C and G bases repeating over and over often occur adjacent to gene-rich areas, forming a barrier between the genes and the "junk DNA." These are called CpG islands and are believed to help regulation of gene activity.

Unlike the human's seemingly random distribution of gene-rich areas, genomes of many other organisms are more uniform, with genes evenly spaced throughout. Humans have on an average three times as many kinds of proteins as the fly or worm because of "alternative splicing" of messenger RNA (mRNA). This process can yield different protein products from the same gene that transcribes mRNA. Humans share most of the same protein families with worms, flies, and plants, but the number of gene family members has expanded in humans, especially in case of proteins involved in development and immunity. The human genome has a much greater portion (50%) of repeat sequences compared to other organisms for e.g., mustard weed (11%), the worm (7%), and the fruit fly (3%).

Scientists have proposed many theories to explain evolutionary contrasts between humans and other organisms, including those of life span, litter sizes, inbreeding, and genetic drift.

3.3 BENEFITS OR APPLICATIONS OF HUMAN GENOME PROJECT

Benefits derived from human genome sequencing are enormous and a few of the applications are mentioned below. There are exceptional opportunities to develop genomic research commercially with production and sale of DNA-based products and technologies that are useful for the following fields.

3.3.1 Molecular Medicine

New era of molecular medicine and biotechnology have emerged from the knowledge derived from HGP. Molecular medicine instead of treating a disorder based on the symptoms, it digs at the root causes of diseases. It aims at developing rapid and more accurate diagnostic tests or genetic screening for early detection of many diseases that enables effective treatment especially for single gene disorders. In addition it looks into genetic factors causing susceptibilities to common complex conditions (like diabetes, hypertension, heart disease, etc.,) in conjecture to environmental conditions and habits/addiction of the persons to smoking etc. Such information will help in assessing the extent of risk and predict the likely onset of a disorder even before it is expressed in an individual. That is it enables "Preclinical" or "Pre-symptomatic" diagnosis by using DNA probes (short stretches of DNA sequences synthesized with base sequence that is complementary to the target gene sequence) that are specifically designed for the detection of different diseases/disorders even before they are expressed. It is also possible to replace the defective genes by the normal genes by the method called "Gene Therapy". In this method the normal gene or a target DNA sequence is incorporated into a vector (bacterial plasmid/a virus/liposome etc.,) and then transferred to patient's tissue grown in culture. Once the target sequence is transfected i.e. incorporated into the recipient cells they are tested for expression

of the transferred sequence or gene and then the tissue is grafted back into the patient where the incorporated normal gene will start functioning and the disease symptoms would disappear. Further, using genome sequence data novel therapeutic regimen can be developed using new classes of drugs, immunotherapy techniques and supplementing with the missing or defective protein.

3.3.2 Risk Assessment

It is understood from human genome analysis that nucleotide differences exist between different individuals which may be associated with their susceptibility or resistance to disease causing factors. Such an information will also be useful in assessing health damage and risks caused by exposure of individuals to radiations including long term low dose exposures and exposure to chemicals and toxins that induce harmful mutations and cancers and infections. This knowledge will help in modulating necessary preventive measures to maintain general health status and healthy society.

3.3.3 Energy and Environment

DOE initiated in 1994 for the Microbial genome Programme to sequence the genomes of bacteria which provide knowledge to benefit human health and environment apart from improving economy from industrial applications. Characterisation of complete microbial genomes will lead to the development of new energy related biotechnologies a) like photosynthetic systems, b) production of biofuels, c) microbial systems that work in extreme environments and also d) organisms that can metabolise readily available renewable resources and waste material. It is possible to develop diverse new products, processes and test methods that would help in maintaining pollution free environment. Above all knowledge of bacterial genomes helps pharmaceutical industries to identif how the pathogenic microbes cause diseases, in detecting new drug targets, identify the minimum number of genes necessary for maintaining life process and stand as models for understanding biological interactions and evolutionary history.

3.3.4 Anthropology, Evolution and Human Migration

Genomic information facilitated a) understanding of human evolution through germ line mutations in lineages b) knowing common biology the humans share with all other life c) study migration pattern of different population groups based on female genetic inheritance d) trace lineage and migration of males through the study of Y chromosome e) identify mutations and compare breakpoints in the evolution of mutations with ages of populations and historical events.

3.3.5 Forensic Science

Genome sequences are species specific and unique to different individuals. Hence genome sequence information used in forensic science is used for a) identifying victims who committed crimes, b) exonerate persons who are wrongly accused, c) identify crime and catastrophic victims d) establish paternity and identify relatives in cases of disputed parentage and e) matching the organ donors with that of recipient for organ transplantation. In addition identification of endangered and protected species among the wild life can be identified by analyzing their genomes of such species. It is possible to detect bacteria and other organisms that may pollute air, water, soil and food. The genomic information also helps in determining pedigrees of plant and livestock in breeding experiments.

3.3.6 Agriculture and Livestock Breeding Drought

Understanding of plants and human genomes allows the creation of disease resistant plants and more nutritious and pesticide free foods. Already the bio-engineered seeds that are insect, pests and drought resistant are being marketed. Similarly disease resistant live stocks and those that are more productive for meat and milk yield are also being developed using genome information.

3.4 DISADVANTAGES OF HGP

The HGP which yielded enormos benefits for scientific research and mankind also led to fears and concerns about the information generated specially about an individual affected with genetic disease for which diagnostic or predictive tests are availble. The major disadvantage is the discrimination by the fellowmen and society which an individual suffers when affected with a genetic disorder. Such individuals are deprived of insurance coverage and will have to face difficulties to meet the medical bills which could be exorbitant. Further they may lose employment opportunities and those employed may be fired by the employers as they fear that an affceted employee may create safety risk at the work place, to the customers and also other employees specially when the genetic condition affects the coordination and judgement as in case of some neurological disorders. While the genetic screening can benefit a family by providing measures for preventing the recurrence among other memebers, it can also destroy the marriages and family relationships. There are also chances of misusing the genomic information by persons with selfish motives and destructive attitude. This will have a tremendous negative effect. In addition to the government, researchers and scientists, people from all walks of life should realise the negative effects and curb them as HGP offers heaps of benefits to the mankind and we have to reap the benefits it offers.

3.5 POST GENOMIC ERA STUDIES

With the availability of genomic sequences from microbes to man, focus is laid on "functional genomics" that provides greater understanding of secrets of life. Research in post genomic era is being focused on:

- Transcriptomics - that involves large-scale analysis of messenger RNAs transcribed from active genes to follow when, where, and under what conditions such genes are expressed.
- Proteomics - that involves study of protein expression and function that explains actual happenings in the cell. This has direct application in designing drugs to treat several genetic diseases/conditions.
- Structural genomics - that generates the 3-D structures of one or more proteins from each protein family that offers clues to function and biological targets for drug designing.
- Experimental methods - for understanding the function of DNA sequences and the proteins they encode including knockout studies to inactivate genes (that are defective or undesirable) in living organisms and monitor changes that could reveal their functions.

- Comparative genomics - that involves analysis of DNA sequence patterns of humans and well studied model organisms for identifying human genes and interpreting their function

3.6 NEED FOR INDIVIDUAL DIPLOID HUMAN GENOME SEQUENCE

Originally HGP aimed at developing haploid reference genome that comprises 3.2 billion nucleotides. Other groups like International HapMap project, Appied Biosystem, Illumina, J. Craig Venter Institute(JCVI), Personal genome Project and Roche undertook the extension of obtaining reference sequence of diploid human genomes. On September 4th, 2007 Craig Venter's complete DNA sequence was published unveiling for the first time the 6 billion nucleotide genome (diploid) of a single individual. His genome was sequenced from the 32 million sequence reads or more than 20 billion base pairs of DNA produced. The diploid genome sequences uniquely catalogued the contributions of the parental chromosomes (in which two sets of chromosomes one from his father and the other from his mother are represented) showing the amount of variation existing between the two. The human reference genome (HuRef) analysis now revealed that:

- The human to human variation is 5-7 times greater as compared to that reported in the earlier haploid genome analysis. This works out to a difference of 15-30 million base pairs between individuals.
- There are 4.1 million DNA variants in an individual of which 22% are non-SNP variants (RFLPs, VNTRs and microsatellites) but they account for about 74% of all the variants found in the DNA.
- There are 3.2 million SNPs and nearly non-SNP variants that include indels (insertion/deletion of nucleotides), copy number variants, block substitutions and segmental duplications. In Venter's genome there were 1.2 million variants that were never before reported.

Analysis of diploid genome generated more informed haplotype assemblies. Haplotypes are linked variations found along the chromosomes (i.e. a set of alleles of different genes located on the same homologue with defined distance). The average occurrence of several haplotypes is reported in populations but not in individuals. Information on individual haplotypes enables study of rare or "private" variants which helps in predicting the traits and diseases in that person. This allows personalised medicine for treating a disease in an individual.

3.7 SPIN OFF OF HGP

3.7.1 1000 Genomes Project

Any two humans are considered to be more than 99 percent the same at the genetic level. However, it is important to understand the small fraction of genetic material that varies among people because it can help explain individual differences causing susceptibility to disease, response to drugs or reaction to environmental factors. To meet this end the "1000 Genomes Project" an international research effort to establish the most detailed catalogue of human genetic variation was launched in January 2008. The project aimed to cover

sequencing the genomes of at least one thousand anonymous participants from a number of different ethnic groups within three years time. With the expertise of multidisciplinary research teams, the 1000 Genomes Project will develop a new map of the human genome that will provide a view of biomedically relevant DNA variations at a greater resolution. The data generated from the 1000 Genomes Project is made swiftly available to the worldwide scientific community through freely accessible public databases. The consortium is expected to generate a valuable tool for all fields of natural science, especially genetics, medicine, pharmacology, biochemistry and bioinformatics.

In 2010, the 1000 genome project finished its pilot phase and increased the sample target to 2000 individuals to be studied by the end of 2010. Still larger project proposed by Wellcome Trust to sequence 10,000 human genomes in three years time to evaluate variantion specially related to diseases. To expand the link of genomic data to observable traits, Church from Harvard Institute, launched "Personal Genome Project" that ultimately aims to sequence 100,000 individuals who voluntarly share their medical records and lifestyle facts. These attempts would generate enoromous information about sequence variations in humans which has lot of applications in treating genetic disaeses, devloping new drugs, population diversity and human evolution.

3.7.2 Haplotype Map or HapMap

One of the projects that emerged as an off shoot of HGP is the "haplotype map" or "HapMap" project which is a tool that allows detection of genes and genetic variations that affect health and disease. The concept of HapMap was based on the view that though any two unrelated persons appear similar in that they share about 99.5% of their DNA sequence, the small fraction of difference between them may greatly affect the risk of an individual to develop a disease. Variation between any two persons are observed to occur at a single nucleotide level i.e. if one has an Adenine (A) nucleotide at a particular site on a chromosome other person may have a Guanine (G) nucleotide at the same position. Such a site is referred to as a single nucleotide polymorphism (SNP), and each of the two possibilities i.e. presence of A or G is called an "allele". Sets of nearby SNPs on the same chromosome are inherited as blocks. This pattern of SNPs on a block is called a "haplotype". While the blocks contain a large number of SNPs, a few SNPs are enough to uniquely identify the haplotypes in a block. The HapMap is a map of these haplotype blocks and the specific SNPs that identify the haplotypes are called "tag SNPs".

The "International HapMap Project" was set up in October 2002 with the collaborations of researchers at academic centers, non-profit biomedical research groups and private companies in Canada, Japan, Nigeria, the United Kingdom, and the United States. The target set for the completion of the project was three years and the information generated by the project is made freely available to researchers around the world through the database. The project was conducted in phases and the complete data obtained in Phase-I were published on 27th October, 2005 and that of Phase II was published in October, 2007 and Phase III dataset was released in spring 2009. The HapMap project focuses only on common SNPs, those occurring with a frequency of at least 1% of the population.

The HapMap is considered as a valuable tool since it facilitates reduction of number of SNPs required to examine in the entire genome for association with a

disease/phenotype from studying the 10 million SNPs to roughly 500,000 tag SNPs. This makes the genome scan approaches easier in detecting regions of interest or with the genes that are linked to diseases much more efficiently. The advantage is that there is no need to study more number of SNPs than necessary and all regions of the genome can be covered. Initially four populations were selected for inclusion in the HapMap: 30 adult-and-both-parents trios from Ibadan, Nigeria (YRI), 30 trios of U.S. residents of northern and western European ancestry (CEU), 44 unrelated individuals from Tokyo, Japan (JPT) and 45 unrelated Han Chinese individuals from Beijing, China (CHB). Although the haplotypes revealed from these populations should be useful for studying many other populations, parallel studies are also foreseen in additional populations in the project.

The HapMap provides a powerful resource for comparing the genetic factors of two groups of people with and without their response to environmental factors, susceptibility to infection and in the effectiveness of and adverse responses to drugs and vaccines. Using just the tag SNPs, researchers are able to find chromosome regions that have different haplotype distributions in the two groups of people, those with or without a disease or response to environment, drugs etc. This helps greatly in the therapeutic management of diseases.

3.7.3 Protein Structure Initiative

To understand how the genes function, we need to know the structure of the proteins produced by them. Such a study referred as "Structural Genomics" is a large scale study that requires several weeks and is also expensive even to determine a single protein structure. The NIH conducts "Protein Structure Initiative" to understand protein structural families, structural folds and the relation of structure to its function. Several companies are also working on this aspect concentrating on the proteins that are medically useful [Pollack, 2000).

3.7.4 Human Epigenome Consortium

Apart from knowing about the genome products (proteins etc), it is also necessary to know when and in which tissue the genes are swithched on or off to start their function or stop it. Such functions are presumed to be controlled by the epigenetic factors. "Epigenetic regulation" refers to regulatories processes that are not mediated by DNA codes but are carried out by mechanisms such as methylation of DNA and histone modification that is presumed to affect the access of transcription mechanisms of DNA, coding for a protein. Epigenetic regulation in clinical disorders is an emerging area of research. A consortium led by Sanger Center from UK, Max Plank Institute for Molecular Genetics from Berlin and a company called Epigenomics has initiated the study of every methylation site within the human genome – a project which could be as large as HGP itself (Hagman, 2000).

3.7.5 Human Genome Diversity Project (HGDP)

Human genome diversity project (HGDP) aims at finding and understanding the diversity and unity of the entire human species. The HGDP was thought of in 1991 by Luigi Luca Cavalli-Sforza, a population geneticist from Stanford University, USA. He and many genetiscists and anthroplogists were already collecting data and samples from several populations around the world mainly

to understand how the human populations are related or differ from each other. These samples stored in different laboratories spread over the world are of immense value and need to be analysed with proper planning. Cavalli-Sforza and his colleagues (1991) state that 'The populations that can tell us the most about our evolutionary past are those that have been isolated for some time, are likely to be linguistically and culturally distinct and are often surrounded by geographic barriers". Such isolated populations are getting rapidly merged with neighbouring groups and the information needed to reconstruct our evolutionary history is being lost. Apart from this, keeping in view the danger of some populations becoming extinct, Cavalli-Sforza and other population geneticists expressed the urgency of implementing the project - HGDP. Finally HGDP was planned in 1993 under the auspices of HUGO with estimated cost of $23-35 millions with a time scale of 5 years for its completion. The project focused on two objectives 1) to trace the evolution and migration of different human populations 2) to identify genes which confer resistance and volunerability to diseases along with the devlopment of traetment modalities and tests required.

The project involved collection, preservation and analysis of human DNA samples from various ethnic groups from around the world specially from small indigenous endangered groups. Blood, skin and hair samples from hundreds of ethnic groups around the world. New tools are used to store genetic information indefinitely by devloping cell lines and DNA segments using polymerase chain reaction (PCR) technique. Any researcher can have access to these samples for the future studies.

From 5000 populations groups across the world the initial documentation from HGDP planning workshops listed 700 target groups. After facing several criticism and scentific debates, HGDP changed its approach indicating that samples should be collected from minority and majority ethnic groups in industrial countries and emphasised that all groups should agree to participate in the project. It also emphasized that selection of the indigeneous groups should depend largely on with which groups the anthropologists have been working or members of the groups who mediate with the study group and the outside world.

Coming to the benefits foreseen by the HGDP project, it enables researche into 1) human origins, 2) migratory and mating patterns, 3) adaptation, 4) disease identification and 5) forensic anthropology. The anthropologists and archeologists are concerned about the origins of human species. Scientists now claim that humans evolved only in Africa, then spread themselves around the world. But there are also possibilities for the simultaneous evolution from several other locations and HGDP may throw light on this claim. The project is also expected to help in measuring the geneological raltionships between the populations by providing information on the ancient migratory patterns like settlements of America and Australia from Asia apart from providing cues about the evolution, dispersal and current distribution of languages. Further comparing the genetic variations in the neighbour hood populations of indigeneous groups, it is possible to understand to what extent these groups are inbred and how long ago these populations have reached the terrotories which are now occupied by them. Maping the "geography" of human genes will be of great value not only to the population geneticists but also to linguists, anthropologists, archeologists and historians. This makes the implementation of HGDP a valid one.

3.8 ETHICAL, LEGAL AND SOCIAL IMPLICATIONS (ELSI)

A project like HGP is expected to be associated with several serious ethical, legal and social implications (ELSI). Hence 3-5% of the budget allocated for HGP was divereted to meet ELSI. Ethical issues are those that raise questions about what is moral and right, legal issues are those that are concerned with the protection of laws and regulations that should be provided and social issues are those that affect the individuals and society at large. These three aspects are interdependent and should be dealt with promptly. Discussions on these aspects emphsised that clear written consent should be obtained from the paricipants after they are explained about the project, pros and consequences and risks if any. The participants should willingly co-operate with the project proceedings and no force should be imposed. Guide lines have been developed taking care to cover several such points.

3.9 SUMMARY

Human Genome project (HGP) is an international initiative implemented in October 1990 to sequence the entire human genome comprising 3.2 billions, base by base with a cost of $3.0 billions within the time frame of 15 years. The project also supported sequencing of several model organisms including that of fruit fly, yeast, mice, bacteria, nematode etc., since the comparative sequence data can help in identifying the new genes and disorders in the human system. The project also supported the development of technologies for high throughput sequencing and of capabilities of computing and storing the sequenced data in to the free data bases like NCBI, Ensembl, GenBank etc. Two different strategies were used for sequencing 1) Strategy employed by the public fundended International human genome consortium where STS regions of each chromosome were shotgunned and the fragments were cloned in to bacterial artificial chromosomes (BACs). Later the sequences from the BAC clones were arranged into longer fragments by joining the overlaps to obtain entire sequence. 2) Strategy proposed by the company Celera Genomics involved shotgunning of the entire genome, developing scaffolds, arranging the overlapping sequences and then assigning them to known chromosomal STS sites. The HGP though discriminates individuals affected with genetic disorders, it offers several benefits like diagnosing diseases, drug designing leading to personalised medicine, assessing the genotypes offering risk or resistance to infections and environmental factors, reducing environmental pollution in developing economically beneficial plants and live stock etc. HGP information culminated in the development of new approaches in the post genomic era like proteomics, transcriptomics etc., which is expected to help in identifying the functions of genes, effect of epigenetic factors in modifying the functions of the genes and phenotypes in establishing biological relationships, in understanding evolutionary process etc., The efforts put by the researchers in this direction led to the development of additional projects like haplotype map (HapMap) project, Human Genome Diversity projects (HGDP). Keeping in view several sensitive issues like discrimination of individuals and population groups, the project also emphasised and allocated the budget to address the ethical, legal and social implications (ELSI). The project in general opened up several avenues for the researchers to answer the questions raised to benefit the humaninty and society at large.

Suggested Reading and References

Robert F. Weaver. 1999. *Molecular Biology.* McGraw-Hill Publishers.

Ricki Lewis. 2003. *Human genetics-Concepts and Applications*. McGraw-Hill Publishers.

Tom Strachan and Andrew P. Read. 2004. *Human Molecular genetics*. Garland Science, 3rd edition.

Hagman M., 12th May 2000. *Mapping a Subtext in our Genetic Book*, *Science 28.*

Hattori et. al. 18th May 2000. *"The DNA Sequence of Human Chromosome 21" Nature* 405.

US Department of energy genome research programmes: genomics.energy.gov.

Pollack A. 4th July, 2000. "The next chapter in the Book of Life: Structural Genomics" *The New York Times*

Shelley D. Smith. 201. *The Human Genome Project 10 Years Out: Reviewing the Past, Present, and Future of Genomic Medicine.* American Society of Human Genetics (ASHG) 60th Annual Meeting.

The human genome- Viewpoint: Science, 16th February 2001: Vol. 291 No.5507 p.1218.

Robert F. Weaver. 1999. *Molecular Biology.* McGraw-Hill Publishers.

Ricki Lewis. 2003. *Human genetics-Concepts and Applications*. McGraw-Hill Publishers.

Tom Strachan and Andrew P. Read. 2004. *Human Molecular genetics*. Garland Science, 3rd edition.

Web sites :

http://www.genome.gov/10001772

http://www.ornl.gov/sci/techresources/Human_Genome/project/info.shtml

http://www.ornl.gov/TechResources/Human_Genome/hg5yp/

http://www.stanford.edu/group/morrinst/HGDP.html

http://www.ornl.gov/sci/techresources/Human_Genome/project/about.shtml.

Sample Questions

1) What is a genome? What do you know about the Human genome Project, its origin, development and implementation?
2) What are the objectives proposed by the Human Genome Project?
3) What were the strategies adopted to sequence the human genome?
4) Enumerate the benefits of the implementation of Human Genome Project.
5) What do you understand by Human Genome Diversity Project? How does it help in understanding the evolution of mankind?
6) Explain briefly the devlopments foreseen in post genomic era.

GLOSSARY

1000 genome project : **1000 Genomes Project**, launched in January 2008, is an international research effort to establish by far the most detailed catalogue of human genetic variation. Scientists plan to sequence the genomes of at least one thousand anonymous participants from a number of different ethnic groups within the next three years, using newly developed technologies which are faster and less expensive.

Alternative Splicing : Various ways of splicing out introns in eukaryotic pre-mRNAs resulting in one gene producing several different mRNAs and protein products.

Bacterial Artificial Chromosome (BAC) : A **bacterial artificial chromosome (BAC)** is an engineered DNA molecule used to clone DNA sequences in bacterial cells (E. coli). Segments of an organism's DNA, ranging from 100,000 to 300,000 bps can be inserted into BACs. The BACs with inserted DNA are then taken up by bacterial cells and as they grow and divide, they amplify the BAC DNA which can then be isolated and used in sequencing.

cDNA : DNA synthesized by reverse transcriptase using RNA as a template.

Copy number variations: are alterations of the DNA of a genome that results in the cell having an abnormal number of copies of one or more sections of the DNA.

CpG island : CpG islands or CG islands are genomic regions that contain a high frequency of CpG sites.

Diploid : The state of having each chromosome in two copies per nucleus or cell. A cell having two chromosome sets, or an individual having two chromosome sets in each of its cells.

Epigenesis : The theory that an individual is developed by successive differentiation of an unstructured egg rather than by a simple enlarging of a preformed entity. The theory holding that development is a gradual process of increasing complexity. For example, organs are formed de novo in the embryo rather than increasing in size from pre-existing structures.

Epigenetic factors : Any factor that is responsible for gene activity/ inactivity without altering the base sequence by way of substitution, insertion or deletion. This factor may alter histones or/ and DNA methylation.

Gene therapy : The correction of a genetic deficiency in a cell by the addition of new DNA and its insertion into the genome.

Genetic mapping : Done based on co-segregation of disease and marker loci & determination of lod scores (likelihood ratios) and is called Linkage Analysis.

Haploid : The state of having one copy of each chromosome per nucleus or cell. A cell having one chromosome set, or an organism composed of such cells.

Haplotype : set of closely linked genetic markers present on one chromosome which tend to be inherited together (not easily separable by recombination).

HGP : The **Human Genome Project** (**HGP**) is an international scientific research project with a primary goal of determining the sequence of chemical base pairs which make up DNA, and of identifying and mapping the approximately 20,000–25,000 genes of the human genome from both a physical and functional standpoint.

In silico analysis : Analysis performed using the computers in conjunction with informatics capabilities.

Linkage : is the tendency of certain loci or alleles to be inherited together. Genetic loci that are physically close to one another on the same chromosome tend to stay together during meiosis, and are thus genetically *linked*.

Methylation : The modification of a strand of DNA after it is replicated, in which a methyl (CH_3) group is added to any cytosine molecule that stands directly before a guanine molecule in the same chain.

mRNA : An RNA molecule transcribed from the DNA of a gene, and from which a protein is translated by the action of ribosomes. The basic function of the nucleotide sequence of mRNA is to determine the amino acid sequence in proteins.

Mutation : Mutation is a permanent change in the DNA sequence of a gene. Mutations in a gene's DNA sequence can alter the amino acid sequence of the protein encoded by the gene.

ORF : A section of a sequenced piece of DNA that begins with an initiation (methionine ATG) codon and ends with a nonsense codon. ORFs all have the potential to encode a protein or polypeptide, however many may not actually do so.

Personal genome project : The **Personal Genome Project** (PGP) is a long term, large cohort study which aims to sequence and publicize the complete genomes and medical records of 100,000 volunteers, in order to enable research into personalised medicine.

Polymorphism : Genetic Polymorphism is the presence of more than two allelic forms at a given locus in such frequencies in a population that the rarest of them is not just due to recurring mutations but is due to a phenomenon called "polymorphisms". The frequency of the rarest allele/form as a rule is taken as > 1.0%.

Pre-symptomatic : Relates to the early phases of a disease when accurate diagnosis is not possible because symptoms of the disease have not yet appeared.

Recombination : is a process by which a molecule of nucleic acid (usually DNA, but can also be RNA) is broken and then joined to a different one. Recombination can occur between similar molecules of DNA, as in homologous recombination, or dissimilar molecules, as in non-homologous end joining.

Regulatory motifs : A sequence motif is a nucleotide or amino acid sequence pattern that is widespread and has, or is conjectured to have, a biological significance.

Scaffold : The eukaryotic chromosome structure remaining when DNA and histones have been removed; made from nonhistone proteins. The central framework of a chromosome to which the DNA solenoid is attached as loops; composed largely of topoisomerase.

Segmental duplication : Segmental duplications are segments of DNA with near-identical sequence.

Sequence tagged site (STS) : Any site in a chromosome or genome that is identified by a known unique DNA sequence. STSs can be used to form genetic maps by standard mapping procedures.

Sequencing : Determination of the order of nucleotides (base sequences) in a DNA or RNA molecule or the order of amino acids in a protein.

Shotgun : Cloning a large population of different DNA fragments, known to contain a fragment of interest, as a prelude to selecting or screening for that one particular clone containing the fragment of interest for intensive study.

Sonication : The process of dispersing, disrupting or inactivating biological material (e.g. viruses) by sound waves.

YAC : A **yeast artificial chromosome** (**YAC**) is a human engineered DNA molecule that acts as vector. Segments of an organism's DNA, ranging one million bps can be inserted into YACs. The YACs, with inserted DNA are then taken up by the yeast cells. As the yeast cells grow and divide, they amplify the YAC DNA, which can then be isolated and used for the physical mapping of complex genomes and for the cloning of large genes.

In the following years researchers attempted to map several other disease genes using different polymorphic loci [See Box-2] related to serum proteins, enzymes and leucocyte (HLA) antigens. In later years DNA or molecular markers like restriction fragment length polymorphisms (RFLPs), variable number of tandem repeats (VNTRs) or minisatellites, 1-4 nucleotides repeats (di, tri and tetra nucleotide repeats) called microsatellites and single nucleotide polymorphisms (SNPs or snips) were discovered and were used in gene mapping studies.