UNIT 3  HUMAN GENOME PROJECT

Contents

3.1  Introduction

3.2  Human Genome Project (HGP)
   3.2.1  History of HGP
   3.2.2  Goals of Human Genome Project
   3.2.3  Strategies of Sequencing
      3.2.3.1  Approach of International Consortium
      3.2.3.2  Approach of Celera Genomics
   3.2.4  Genome Donors for Sequencing
   3.2.5  Genome Assembly
   3.2.6  Genome Annotation
   3.2.7  Observations Drawn from Human Genome Sequencing
   3.2.8  How is the Human Genome Arranged or Organised?

3.3  Benefits or Applications of Human Genome Project
   3.3.1  Molecular Medicine
   3.3.2  Risk Assessment
   3.3.3  Energy and Environment
   3.3.4  Anthropology, Evolution and Human Migration
   3.3.5  Forensic Science
   3.3.6  Agriculture and Livestock Breeding Drought

3.4  Disadvantage of HGP

3.5  Post Genomic Era Studies

3.6  Need for Individual Diploid Human Genome Sequence

3.7  Spin Off of HGP
   3.7.1  1000 Genomes Project
   3.7.2  Haplotype map or HapMap
   3.7.3  Protein Structure Initiative
   3.7.4  Human Epigenome Consortium
   3.7.5  Human Genome Diversity Project (HGDP)

3.8  Ethical, Legal and Social Implications (ELSI)

3.9  Summary
   Suggested Reading and References
   Sample Questions
   Glossary

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

<table>
<thead>
<tr>
<th><em>Gene with known sequence</em></th>
<th>Autosomal</th>
<th>X-Linked</th>
<th>Y-Linked</th>
<th>Mitochondrial</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>13017</td>
<td>638</td>
<td>48</td>
<td>35</td>
<td></td>
<td>13738</td>
</tr>
<tr>
<td>+ Gene with known sequence and phenotype</td>
<td>171</td>
<td>6</td>
<td>0</td>
<td>2</td>
<td>179</td>
</tr>
<tr>
<td># Phenotype description, molecular basis known</td>
<td>3048</td>
<td>258</td>
<td>4</td>
<td>28</td>
<td>3338</td>
</tr>
<tr>
<td>% Mendelian phenotype or locus, molecular basis unknown</td>
<td>1654</td>
<td>136</td>
<td>5</td>
<td>0</td>
<td>1795</td>
</tr>
<tr>
<td>Other, mainly phenotypes with suspected mendelian basis</td>
<td>1800</td>
<td>129</td>
<td>2</td>
<td>0</td>
<td>1931</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>19690</strong></td>
<td><strong>1167</strong></td>
<td><strong>59</strong></td>
<td><strong>65</strong></td>
<td><strong>20981</strong></td>
</tr>
</tbody>
</table>

#### Table 3.1b: Synopsis of human gene map

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

[Diagram showing parental combinations and recombinants]

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 \( \theta \) 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 (\( H_0 = \text{No linkage}; \theta = 0.5 \)) and comparing this with the probabilities based on a range of recombination fractions \( \theta \) varying from 0.00 to 0.05, i.e.

\[
P_R = \frac{P(\text{FAMILY, GIVEN } \theta = 0.0 \text{ to } 0.5)}{P(\text{FAMILY, GIVEN } \theta = 0.0 \text{ to } 0.5)}
\]

\( P_R \) 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](image)

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 tetrnucleotide 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 microsatellite 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

<table>
<thead>
<tr>
<th>Type of Marker</th>
<th>No. of Loci</th>
<th>Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>Blood Group antigens</td>
<td>&gt;25</td>
<td>Genotype cannot always be inferred from the phenotypes. Difficult for physical localization of genes</td>
</tr>
<tr>
<td>Serum Proteins</td>
<td>30</td>
<td>Often limited polymorphisms. Difficult for physical localization of genes</td>
</tr>
<tr>
<td>Leucocyte antigens (HLA)</td>
<td>1</td>
<td>HLA system with A,B,C,D and DR loci each harbors hundreds of alleles thus resulting extensive polymorphism. Is highly informative useful in gene mapping</td>
</tr>
<tr>
<td>Restriction Fragment Length Polymorphisms (RFLPs)</td>
<td>$&gt;10^5$</td>
<td>Bi-allelic markers, maximum heterozygosity 0.5, genotyped using Southern blotting and PCR techniques. Easy for physical localisation</td>
</tr>
<tr>
<td>Variable number of tandemn repeats (VNTRs)</td>
<td>$&gt;10^4$</td>
<td>Many alleles at each locus. Highly informative. Easy for physical localization of genes. Tend to cluster near the ends of chromosomes</td>
</tr>
<tr>
<td>Microsatellites (di, tri and tetranucleotide repeats)</td>
<td>$&gt;10^4$</td>
<td>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.</td>
</tr>
<tr>
<td>Single Nucleotide Polymorphisms (SNPs)</td>
<td>$4\times10^5$</td>
<td>Shows variation between any two individuals at single nucleotide level. Are most stable but less informative than microsattellites. Can be genotyped on a very large scale using automated sequencing. Useful for genome wide screening, gene mapping and risk prediction for diseases</td>
</tr>
</tbody>
</table>
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 report 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 facilities for human genetic research
1) 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
Human Genome Project

1) Primer Extension Reaction (For the synthesis of new strands using the template sequence)

Template

ACGGTAT

Replication with ddCTP

18 bp primer

ACGGTAT

TG

(21 bases)

2) Four reactions terminated with ddA, ddT, ddC and ddG resulting 4 in products of different sizes

Products of ddA reaction

Template: ACGGTAT

(23) TGCCCA

(25) TGCCATA

Products of ddT reaction

Template: ACGGTAT

(19) T

(24) TGCCAT

Products of ddC reaction

Template: ACGGTAT

(21) TG

(22) TGC

(22) TGCC

Products of ddG reaction

Template: ACGGTAT

(20) T

(19) G

3) Electrophoresis of the synthesized products

<table>
<thead>
<tr>
<th>ddA</th>
<th>ddT</th>
<th>ddC</th>
<th>ddG</th>
</tr>
</thead>
<tbody>
<tr>
<td>A 25</td>
<td>T 24</td>
<td>A 23</td>
<td>C 22</td>
</tr>
<tr>
<td>C 21</td>
<td>G 20</td>
<td>T 19</td>
<td></td>
</tr>
</tbody>
</table>

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 (TGCCCGT) 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.

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.

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

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.
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 undertaken 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 database, sequences of known and hypothetical genes and proteins can be retrieved. The databases are open for everyone 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 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 underlie 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 identify 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 available. 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 affected 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.
Human Genome Project

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, Applied 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 variation 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 voluntarily share their medical records and lifestyle facts. These attempts would generate enormous information about sequence variations in humans which has lot of applications in treating genetic diseases, developing 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 switched 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 regulatory 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 geneticists and anthropologists 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 development of treatment 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 developing 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 scientific 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 indigenous 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 research 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 genealogical relationships 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 neighbourhood populations of indigenous groups, it is possible to understand to what extent these groups are inbred and how long ago these populations have reached the territories which are now occupied by them. Mapping 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 diverted 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 emphasized that clear written consent should be obtained from the participants after they are explained about the project, pros and consequences and risks if any. The participants should willingly cooperate with the project proceedings and no force should be imposed. Guidelines 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 funded 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 personalized medicine, assessing the genotypes offering risk or resistance to infections and environmental factors, reducing environmental pollution in developing economically beneficial plants and livestock 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 emphasized 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 humanity and society at large.
Suggested Reading and References


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


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 developments foreseen in post genomic era.
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.
<table>
<thead>
<tr>
<th><strong>Gene therapy</strong></th>
<th>The correction of a genetic deficiency in a cell by the addition of new DNA and its insertion into the genome.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Genetic mapping</strong></td>
<td>Done based on co-segregation of disease and marker loci &amp; determination of lod scores (likelihood ratios) and is called Linkage Analysis.</td>
</tr>
<tr>
<td><strong>Haploid</strong></td>
<td>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.</td>
</tr>
<tr>
<td><strong>Haplotype</strong></td>
<td>set of closely linked genetic markers present on one chromosome which tend to be inherited together (not easily separable by recombination).</td>
</tr>
<tr>
<td><strong>HGP</strong></td>
<td>The <strong>Human Genome Project (HGP)</strong> 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.</td>
</tr>
<tr>
<td><strong>In silico analysis</strong></td>
<td>Analysis performed using the computers in conjunction with informatics capabilities.</td>
</tr>
<tr>
<td><strong>Linkage</strong></td>
<td>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.</td>
</tr>
<tr>
<td><strong>Methylation</strong></td>
<td>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.</td>
</tr>
<tr>
<td><strong>mRNA</strong></td>
<td>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.</td>
</tr>
<tr>
<td><strong>Mutation</strong></td>
<td>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.</td>
</tr>
<tr>
<td><strong>ORF</strong></td>
<td>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.</td>
</tr>
</tbody>
</table>
**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.