Human Genome Project

Introduction to HGP

- The **Human Genome Project (HGP)** was an international scientific research project that aimed to determine the **complete sequence of nucleotide base pairs** that make up human DNA and all the genes it contains.
- It remains the world's largest collaborative biological project.
- The idea was picked up in 1984 by the **US government** when the planning started, the project was formally launched in **1990** and was declared complete in **2003**.

Introduction to HGP

- The Human Genome Project originally aimed to map the nucleotides contained in a **human haploid reference genome.**
- The "genome" of any given individual is unique; mapping the "human genome" involved **sequencing** the genomes of a small number of individuals and then **assembling** these together to get a complete sequence for each chromosome.
- The finished human genome is thus a **mosaic**, not representing any one individual.

GOALS OF HGP

- To identify and map all the **20,000-25,000 genes** (approx) in the human DNA from a physical and functional standpoint.
- To determine the sequences of the **3 billion chemical base pairs** that make up the human DNA.
- To store these informations in **databases**.
- To discover more **efficient technologies** for data analysis.
- Allow the private sector **access to the informations** and technologies that arise from this project.
- Also to **sequence the genomes of other organisms** that are important in medical research such as mouse, Drosophila etc,.
- To address ethical, legal and social issues.

Participating countries and funding agencies

- In **1990**, the 2 major funding agencies, the **US Department of Energy (DOE) and National Institute of Health (NIH),** developed an MoU in order to coordinate plans and set the clock for the initiation of the Project.
- Most of the government-sponsored sequencing was performed in 20 universities and research centers in the **United States, the United Kingdom, Japan, France, Germany, Canada, and China.**
- A parallel project was conducted outside the government sponsorship by the **Celera Corporation** or the Celera Genomics which was formally launched in 1998.

Participating countries and funding agencies

- The **$3-billion project** was formally launched in 1990 by the US Department of Energy and the National Institute of Health.
- The Human Genome Project was a **13-year-long**, publicly funded project initiated in 1990 with the objective of determining the DNA sequence of the entire **euchromatic human genome** within 15 years.

Pioneers in HGP

- **Robert Sinsheimer** proposed the idea of sequencing the human genome in the year 1985.
- **Charles DeLisi** and **David Smith** proposed the budget for Human Genome Project.
- HGP act was passed in the US congress under **President Regan** in 1988.
- **James Watson** headed the NIH Genome Program.
- **Francis Collins** succeeded James Watson in 1993 as the overall Project Head and the Director of the NIH (which later become the National Human Genome Research Institute **NHGRI**) and was in power until the completion of HGP in 2003.

Pioneers in HGP

- **Jim Kent**, a PhD scholar in the University of California Santa Cruz (UCSC), in May 2000, developed a software, **GigAssembler**, that allowed the publicly funded Human Genome Project to assemble and publish the human genome sequence.

TIMELINE OF HGP

- **1970** – Fredrick Sanger developed a technique for DNA sequencing, known as the Sanger's method of DNA sequencing.
- **1985** - Robert Sinsheimer at UCSC proposed the idea of sequencing the human genome.
- **1986** - the U.S. Dept of Energy and the National Institute of Health came forward to fund the Human Genome Project.

1989 - U.K's medical research council (MRC) joined the Human Genome Project.

TIMELINE OF HGP

1990 – HGP was officially launched with James Watson as its Project Director.
the 1st gene to be mapped was BRCA1, which is the gene for breast cancer.

1993 - 1st 5 year plan for HGP was published.
Sanger Institute(UK) joins HGP.

1994 – HGP's Human genetic mapping goal was achieved.

1995 - Genetic privacy act was passed.
1st bacterial genome was sequenced (Hemophilus influenzae)

1996 – 1st Human Gene map was published.
Yeast genome was sequenced.
HGP's mouse genetic mapping goal was achieved.

TIMELINE OF HGP

1997 - NIH becomes NHGRI.
E.coli genome sequenced.
Genoscope, French National Genome Sequencing Centre was established.

1998 - 2nd 5 year plan for HGP was published.
Japan's RIKEN Genomic Services Centre was established.
Genome of the roundworm Caenorhabditis elegans was sequenced.
SNP sequencing was initiated.
the Chinese National Human Genome Centres were established in Beijing and Shanghai.

1999 - sequencing of human chromosome 22 was completed and was published in "The Nature."

Timeline of HGP

2000 - working draft of human genome completed.
US president Clinton & UK's PM Blair support free access to genome information.
Genomes of D.melanogaster and A.thaliana were sequenced & published in "The Nature".

2001 – working draft of human genome sequence was published in "The Nature" & "Science".

2002 – working draft of mouse genome sequence was completed & published.

2003 - finished version of human genome sequence was completed.
HGP ended with all the goals achieved.

Technical aspects in HGP

- The process of determining the human genome first involves **genome mapping**, or characterizing the chromosomes. This is called a **genetic map.**
- The next step is **DNA sequencing ,**or determining the order of DNA bases on a chromosome. These are **physical maps.**

STRATEGIES

- **Genetic markers** are invaluable for genome mapping.
- Markers are any inherited physical or molecular characteristics that are different among individuals of a population (polymorphic)
- A **genetic map** shows the relative locations of these specific markers on the chromosomes.
- An example of a marker includes restriction fragment length polymorphisms **(RFLP).**
- Used in RFLP markers are restriction enzymes. These enzymes recognize short sequences of DNA and cut them at specific sites, therefore, DNA can be cut into many different fragments. These fragments are the DNA pieces used in physical maps
- RFLPs reflect sequence differences in DNA sites which are cleaved by restriction enzymes.

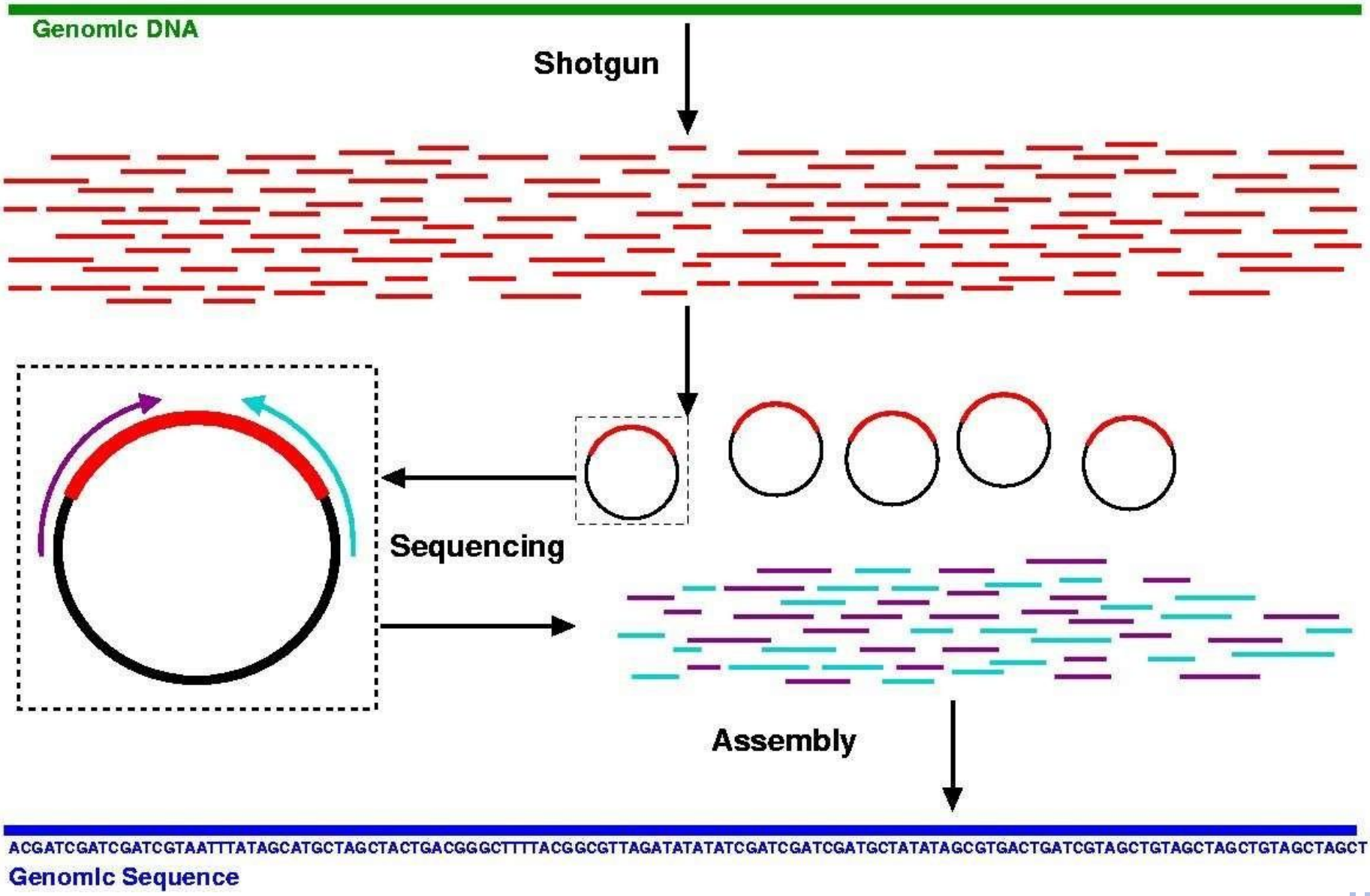
Sequencing strategies

- To sequence DNA, it must be first be **amplified**, or increased in quantity.
- Two types of DNA amplifications are **cloning** and Polymerase Chain Reactions **(PCR).**
- Now that the DNA has been amplified, sequencing can begin.
- Sequencing techniques used in HGP are:-

  1)Shotgun sequencing method &

  2)Sanger sequencing method

SHOTGUN METHOD



SANGER METHOD

Sanger sequencing

- Low-throughput, but accurate and can handle up to 1000bp
- Still standard for small-scale laboratory use

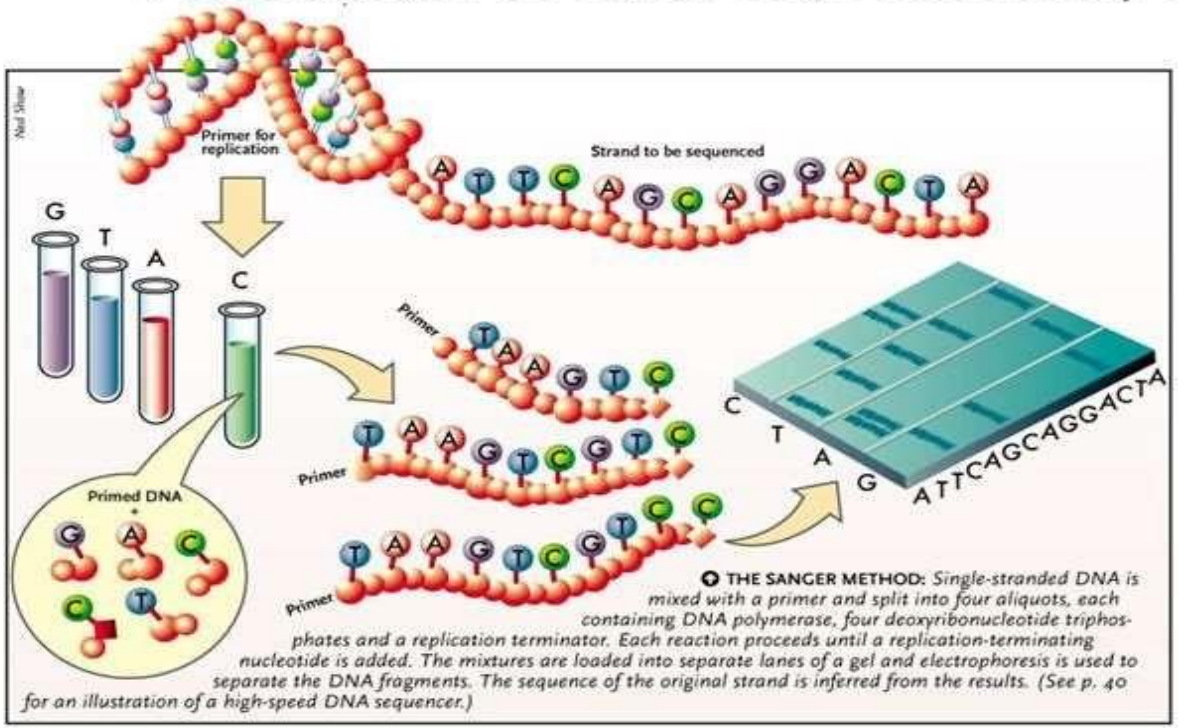


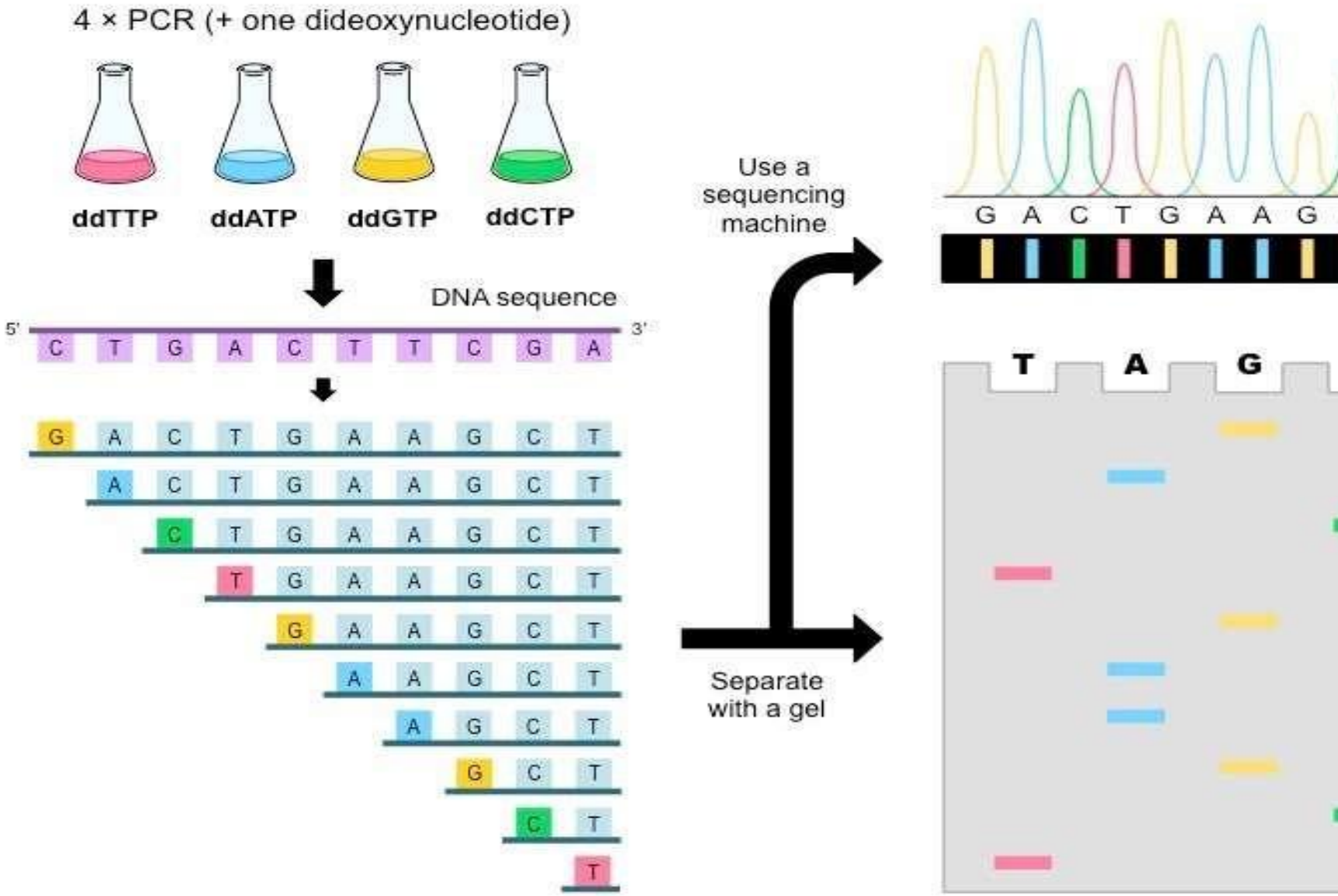
Image credit: the-scientist.com

Components:

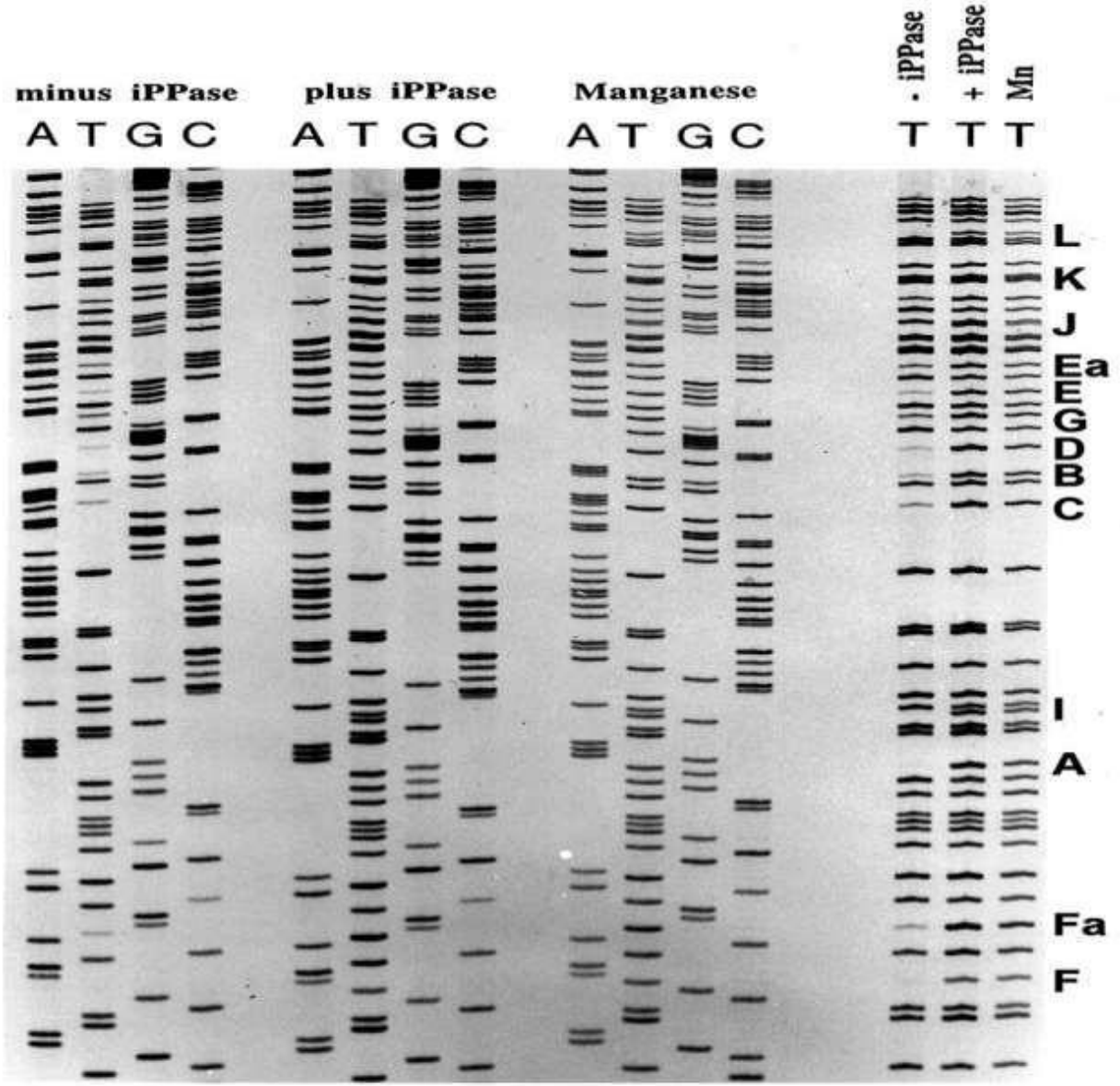
- DNA to be sequenced
- Primer
- Free nucleotides that allow further extension (dNTP, circles):
  - N=A, C, G or T, all four types are present
- Free nucleotides that terminate extension (ddNTP, rhombuses):
  - N=A, C, G or T, only one type is present
- DNA polymerase

See these videos for animations:
http://www.youtube.com/watch?v=oYpIlbI0qF8
http://www.youtube.com/watch?v=6ldtdWjDwes

SANGER SEQUENCING



A



B

5′ TCGAATTCGTAATCATGGTCATAGCTGTTTCCTGTGTGAAATTGTTATCCGCTCACAATT
1 2,3 F 3a Fa 4 + 5 6 (7) 8 8a A 9,10 I 11 12 + 13 14,15

CCACACAACATACGAGCCGGAAGCATAAAGTGTAAAGCCTGGGGTGCCTAATGAGTGAG
+ 16 C 17 B D 18 G 19 E

CTAACTCACATTAATTGCGTTGCGCTCACTGCCCGCTTTCCAGTCGGGAAACCTGTCGTG
20 Ea 21,22 22a,b J 30 23 23a 24 K 25 26 27 28 L

Outcomes of HGP

- There are approximately **22,300** protein-coding genes in human beings, the same range as in other mammals. Mouse – 23,000 genes (approx)
- **Drosophila – 17,000 genes (approx),**
  C.elegans - < 22,000 genes
- we share many homologous genes (called "orthologs") with both these animals. But:-
- many of our protein-encoding genes produce more than one protein product (e.g., by alternative splicing of the primary transcript of the gene). On average, each of our ORFs produces 2 to 3 different proteins.
- So the human "proteome" (our total number of proteins) may be 10 or more times larger than that of the fruit fly and roundworm.

- A larger proportion of our genome :-
  encodes transcription factors
  is dedicated to control elements (e.g., enhancers) to which these transcription factors bind
  The combinatorial use of these elements provides much greater flexibility of gene expression than is found in Drosophila and C.elegans.

Outcomes of HGP

- **Gene density :-**
  23 genes per million base pairs on chromosome 19
  5 genes per million base pairs on chromosome 13.
- Humans, and presumably most vertebrates, have genes not found in invertebrate animals like Drosophila and C. elegans. Few of those genes are :-
  antibodies and T cell receptors for antigen (TCRs)
  the transplantation antigens of the major histocompatibility complex (MHC) & human leucocyte antigen (HLA).
  cell-signaling molecules including the many types of cytokines
  the molecules that participate in blood clotting.
- Human genome comprises of 2% of exons (coding regions) and 98% of introns (non-coding regions).

APPLICATIONS OF HGP

- The sequencing of the human genome holds benefits for many fields, from **molecular medicine** to human evolution.
- Helps in identifying **disease causing gene**.
- identification of **mutations** linked to different forms of cancer.
- The sequence of the DNA is stored in **databases** available to anyone on the Internet.
- The U.S. National Center for Biotechnology Information (and sister organizations in Europe and Japan) house the gene sequence in a database known as **GenBank**, along with sequences of known and hypothetical genes and proteins.
- will allow for advances in agriculture through **genetic modification** to yield healthier, more disease-resistant crops.
- Benefitted the advancement of **forensic science**.