International Journal
of
Pharmacy Review & Research
www.ijprr.com

e-ISSN: 2248 – 9207
Print ISSN: 2248 – 9193

THE HUMAN GENOME PROJECT – A REVIEW

***K. Thamizhvanan, R. Himabindu, K. Sarada, N. Hyndavi, K. Roopa Lahari**

*Department of Pharmaceutical Biotechnology, Sree Vidyanikethan College of Pharmacy,
Sree Sainath Nagar, Chandragiri (M), Tirupati, Andhra Pradesh, India-517102.

ABSTRACT

The objective of the Human Genome Project is to understand the genetic makeup of the human species, the project has also focused on several other nonhuman organisms such as *E. coli*, the fruit fly, and the laboratory mouse. It remains one of the largest single investigative projects in modern science. The mapping of human genes is an important step in the development of medicines and other aspects of health care. The project that has identified and located all of the genes in human DNA, and determined the sequences of the chemical bases that make up human DNA. This information is stored in databases. A four billion-dollar program known as the Human Genome Project has been center of attention in scientific research.

Keywords: Human Genome Project, DNA extraction, detecting DNA, Applications.

INTRODUCTION

The project that has identified and located all of the genes in human DNA, and determined the sequences of the chemical bases that make up human DNA. This information is stored in databases. A four billion-dollar program known as the Human Genome Project has been center of attention in scientific research. 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.. An international collaboration to map the genetic blue print of a human is the main focus.

The Human Genome Project has grabbed the attention of medical professionals, scientists, insurance companies and also every other individual out there concerned with his health background. Hopefully, one will become better informed and decide to spread the knowledge so ignorance does not hurt what could be a revolutionary medical answer. The objective of the Human Genome Project is to understand the genetic makeup of the human species, the project has also focused on several other nonhuman organisms such as *E. coli*, the fruit fly, and the laboratory mouse. It remains one of the largest single investigative projects in modern science. The mapping of human genes is an important step in the development of medicines and other aspects of health care [1].

HGP arose from two key insights in the early 1980s.

- The ability to take global views of genomes could greatly accelerate biomedical research by allowing researchers to attack problems in a comprehensive fashion.
- The creation of such global views would requires a communal effort in infrastructure research.

Key projects helped to crystallize the insights, including

i) The sequencing ofthe some bacterial and animal viruses, as well as the human
mitochondrion between 1977 and 1982.
ii) The development of (random) shotgun sequencing of long DNA fragments for high
throughput gene discovery, later dubbed with expressed sequence tags(ETSs) and
assembling computer programs.

Corresponding Author **K.Thamizhvanan E mail:** ktvanan2006@Yahoo.co.in

The main reasoning behind this program is to better understand diseases and how the body works. Preventive strategies will be the tool of medicine in the near future. Instead of cures for sicknesses after the fact, people can utilize new preventive strategies to avoid illness [2].

Human Genome Project base :

The Human Genome Project (HGP) is an international research program designed to:

1. Construct detailed genetic and physical maps of the human genome,
2. Determine the complete nucleotide sequence of human DNA,
3. To localize the 'estimated' 50,000 – 100,000 genes within the human genome,
4. Perform similar analyses on the genomes of other organisms used as model systems in research.
5. Produce a resource of detailed information about the structure organization and function of human DNA, information that constitutes the basic set of inherited "instructions" for the development and functioning of a human being.

Objectives

- Determine the sequences that comprise human DNA,
- Identify all of the genes in human DNA,
- Store this information in databases and improve tools available for its analysis,
- Transfer technologies gained from the project to private industry (eg biotechnology companies) to develop new medical applications,
- Address the ethical, legal and social issues that may arise from the project.
- The main aims of the Human Genome Project (HGP) were to:
- Construct maps of the genome (genetic and physical)
- Identify all the genes (now known to be about 30,000)
- Determine the entire DNA sequence (3,000,000,000 bp)
- Genetic mapping of the human genome – ie locating 3000 genetic markers n human DNA (genes)
- Physical mapping – ie cutting each chromosome into fragments and then determining the correct order of the pieces;
- DNA sequencing – determining the exact order of the nucleotides on each chromosome.
- Analysing the genomes of other organisms eg bacteria, yeast [3].

GENES

- For every 2 biologists, you get 3 definitions"A DNA sequence that encodes a heritable trait."
- The unit of heredity
- "Classic" vs. "modern" understanding of molecular biology

Genome Confusion

The sequence of a gene in the genome includes:

- protein coding sequence
- introns and exons
- 5' and 3' untranslated regions on the mRNA
- promoter and 5' transcription factor binding sites
- enhancers
- Multiple cDNAs with different sequences (that produce different proteins) can be transcribed from the same genomic locus

Finding genes in genome sequence is not easy

- About 1% of human DNA encodes functional genes.
- Genes are interspersed among long stretches of non-coding DNA.
- Repeats, pseudo-genes, and introns confound matters
- The current estimate is 34,000 human genes.
- The same number as the mouse, only about 5 times more than yeast [4].

Yet two different versions of the human genome (Celera vs. Ensembl/UCSC) show only about 50% overlap between the genes that they have described.

Data Mining Tools

- Scientists need to work with a lot of layers of information about the genome
- coding sequence of known genes and cDNAs
- genetic maps (known mutations and markers)
- gene expression
- cross species homolog

DNA EXTRACTION:

- Breaking the cells open, commonly referred to as cell disruption or cell lysis, to expose the DNA within. This is commonly achieved by grinding or sonicating the sample.
- Removing membrane lipids by adding a detergent.
- Removing proteins by adding a protease (optional but almost always done).
- Removing RNA by adding an RNase (often done).
- Precipitating the DNA with an alcohol — usually ice-cold ethanol or isopropanol. Since DNA is insoluble in these alcohols, it will aggregate together, giving a *pellet* upon centrifugation. This step also removes alcohol-soluble salt.
- Refinements of the technique include adding a chelating agent to sequester divalent cations such as Mg^{2+} and Ca^{2+}, which prevents enzymes like DNase from degrading the DNA.
- Cellular and histone proteins bound to the DNA can be removed either by adding a protease or by having precipitated the proteins with sodium or ammonium acetate, or extracted them with a phenol-chloroform mixture prior to the DNA-precipitation.
- If desired, the DNA can be resolubilized in a slightly alkaline buffer or in ultra-pure water [5].

Detecting DNA

A diphenylamine (DPA) indicator will confirm the presence of DNA. This procedure involves chemical hydrolysis of DNA: when heated (e.g. ≥95 °C) in acid, the reaction requires a deoxyribose sugar and therefore is specific for DNA. Under these conditions, the 2-deoxyribose is converted to w-hydroxylevulinyl aldehyde, which reacts with the compound, diphenylamine, to produce a blue-colored compound. DNA concentration can be determined measuring the intensity of absorbance of the solution at the 600 nm with a spectrophotometer and comparing to a standard curve of known DNA concentrations.

Measuring the intensity of absorbance of the DNA solution at wavelengths 260 nm and 280 nm is used as a measure of DNA purity. DNA absorbs UV light at 260 and 280 nanometres, and aromatic proteins absorb UV light at 280 nm; a pure sample of DNA has the 260/280 ratio at 1.8 and is relatively free from protein contamination. A DNA preparation that is contaminated with protein will have a 260/280 ratio lower than 1.8. DNA can be quantified by cutting the DNA with a restriction enzyme, running it on an agarose gel, staining with ethidium bromide or a different stain and comparing the intensity of the DNA with a DNA marker of known concentration.

Using the Southern blot technique, this quantified DNA can be isolated and examined further using PCR and RFLP analysis. These procedures allow differentiation of the repeated sequences within the genome. It is these techniques which forensic scientists use for comparison, identification, and analysis [6].

Instrumentation used in DNA Extraction

DNA FRAGMENTATION

- DNA-modifying enzymes--Nucleases and ligases
- Nucleases are enzymes that cut DNA strands by catalyzing the hydrolysis of the phosphodiester bonds.
- Nucleases that hydrolyse nucleotides from the ends of DNA strands are called exonucleases, while endonucleases cut within strands.
- The most frequently used nucleases in molecular biology are the restriction endonucleases, which cut DNA at specific sequences.
- For instance, the EcoRV enzyme shown to the left recognizes the 6-base sequence 5′-GAT|ATC-3′ and makes a cut at the vertical line.
- In nature, these enzymes protect bacteria against phage infection by digesting the phage DNA when it enters the bacterial cell, acting as part of the restriction modification system.
- In technology, these sequence-specific nucleases are used in molecular cloning and DNA fingerprinting.
- Enzymes called DNA ligases can rejoin cut or broken DNA strands.
- Ligases are particularly important in lagging strand DNA replication, as they join together the short segments of DNA produced at the replication fork into a complete copy of the DNA template.
- They are also used in DNA repair and genetic recombination [7].

DNA ISOLATION

DNA isolation is a routine procedure to collect DNA for subsequent molecular or forensic analysis.

A. Large scale double-stranded DNA isolation

The method used for the isolation of large scale cosmid and plasmid DNA is an unpublished modification (16) of an alkaline lysis procedure (17,18) followed by equilibrium ultracentrifugation in cesium chloride-ethidium bromide gradients (1). Briefly, cells containing the desired plasmid or cosmid are harvested by centrifugation, incubated in a lysozyme buffer, and treated with alkaline detergent. Detergent solubilized proteins and membranes are precipitated with sodium acetate, and the lysate is cleared first by filtration of precipitate through cheesecloth and then by centrifugation. The DNA-containing supernatant is transferred to a new tube, and the plasmid or cosmid DNA is precipitated by the addition of polyethylene glycol and collected by centrifugation. The DNA pellet is resuspended in a buffer containing cesium chloride and ethidium bromide, which is loaded into polyallomer tubes and subjected to ultracentrifugation overnight. The ethidium bromide stained plasmid or cosmid DNA bands, equilibrated within the cesium chloride density gradient after ultracentrifugation, are visualized under long wave UV light and the lower band is removed with a 5 cc syringe. The intercalating ethidium bromide is separated from the DNA by loading the solution onto an equilibrated ion exchange column. The A260 containing fractions are pooled, diluted, and ethanol precipitated, and the final DNA pellet is resuspended in buffer and assayed by restriction digestion as detected on agarose gel electrophoresis.

During the course of this work several modifications to the above protocol were made. For example, initially cell growth times included three successive overnight incubations, beginning with the initial inoculation of 3 ml of antibiotic containing media with the plasmid or cosmid-containing bacterial colony, and then increasing the culture volume to 50 ml, and then to 4 l. However, it was observed that recombinant cosmid DNA isolated from cell cultures grown under these conditions, in contrast to recombinant plasmid DNA, was contaminated with deleted cosmid DNA molecules. However, these deletions are avoided by performing each of the three successive incubations for eight hours instead of overnight, although a slight yield loss accompanied the reduced growth times [8].

Recently, a diatomaceous earth-based (19-22) method was used to isolate the plasmid or cosmid DNA from a cell lysate. The cell growth, lysis, and cleared lysate steps are performed as described above, but following DNA precipitation by polyethylene glycol, the DNA pellet is resuspended in RNase buffer and treated with RNase A and T1. Nuclease treatment is necessary to remove the RNA by digestion since RNA competes with the DNA for binding to the diatomaceous earth. After RNase treatment, the DNA containing supernatant is bound to diatomaceous earth in a chaotropic buffer of guanidine hydrochloride by incubation at room temperature. The DNA-associated diatomaceous earth then is collected by centrifugation, washed several times with ethanol buffer and acetone, dried, and then resuspended in buffer. The DNA is eluted during incubation at 65degC, and the DNA-containing supernatant is collected after centrifugation and separation of the diatomaceous earth particles. The DNA recovery is measured by taking absorbance readings at 260 nanometers. After concentration by ethanol precipitation, the DNA is assayed by restriction digestion [9].

Protocol

1. Pick a colony of bacteria harboring the plasmid or cosmid DNA of interest into a 12 X 75 mm Falcon tube containing 2 ml of LB media supplemented with the appropriate antibiotic (typically ampicillin at 100 ug/ml) and incubate at 37deg C 8-10 hours with shaking at 250 rpm. Transfer the culture to an Ehrlenmeyer flask containing 50 ml of similar media, and incubate further for 8-10 hours. Transfer 12.5 ml of the culture to each of 4 liters of similar media, and incubate for an additional 8-10 hours.
2. Harvest the cells by centrifugation at 7000 rpm for 20 minutes in 500 ml bottles in the RC5-B using the GS3 rotor. Resuspend the cell pellets in old media and transfer to two bottles, centrifuge as before, and decant the media. The cell pellets can be frozen at -70degC at this point.
3. Resuspend the cell pellets in a total of 70 ml of GET/Lysozyme solution (35 ml for each bottle) by gently teasing the pellet with a spatula and incubate for 10 minutes at room temperature. (Note: Do not vortex the lysate at any time because this may shear the chromosomal DNA).
4. Add a total of 140 ml of alkaline lysis solution (70 ml for each bottle), gently mix, and incubate for 5 minutes in an ice-water bath.
5. Add 105 ml of 3M NaOAc, pH 4.8 (52.5 ml for each bottle), cap tightly, gently mix by inverting the bottle a few times, and incubate in an ice-water bath for 30-60 minutes.
6. Clear the lysate of precipitated SDS, proteins, membranes, and chromosomal DNA by pouring through a double-layer of cheesecloth. Transfer the lysate into 250 ml centrifuge bottle, centrifuge at 10,000 rpm for 30 minutes at 4deg C in the RC5-B using the GSA rotor [10].

For cesium chloride-gradient purification:

7. Pool the cleared supernatants into to a clean beaker, add one-fourth volume of 50% PEG/0.5 M NaCl, swirl to mix, and incubate in an ice-water bath for 1-2 hours.
8. Collect the PEG-precipitated DNA by centrifugation in 250 ml bottles at 7000 rpm for 20 minutes at 4degC in the RC5-B using the GSA rotor.
9. Dissolve the pellets in a combined total of 32 ml of 100:10 TE buffer, 5 ml of 5 mg/ml ethidium bromide, and 37 g cesium chloride (Var Lac Oid Chemical Co., Inc.) (final concentration of cesium chloride should be 1 g/ml).
10. Transfer the sample into 35 ml polyallomer centrifuge tubes, remove air bubbles, seal with rubber stoppers, and crimp properly.
11. Centrifuge at 60,000 rpm to 16-20 hours at 15-20degC in the Sorvall OTD-75B ultracentrifuge (DuPont) using the T-865 rotor.
12. Visualize the ethidium bromide stained DNA under long-wave UV light, and remove the lower DNA band using a 5 cc syringe and a 25 gauge needle. (It may be helpful first to remove and discard the upper band).
13. To remove the ethidium bromide, load the DNA sample onto an equilibrated 1.5 ml Dowex column, and collect 0.5 ml fractions. Equilibrate the Dowex AG resin (BioRad) by successive centrifugation, resuspension, and decanting with 1M NaOH, water, and then 1M Tris-HCl, pH 7.6 until the Dowex solution has a pH of 7.6.
14. Pool fractions with an A260 of 1.00 or greater into 35 ml Corex glass tubes, add one volume of ddH2O, and ethanol precipitate by adding 2.5 volumes of cold 95% ethanol. Incubate at least 2 hours at -20degC, centrifuge at 10,000 rpm for 45 minutes in the RC5-B using the SS-34 rotor. Gently decant the supernatant, add 80% ethanol, centrifuge as before, decant, and dry the DNA pellet in a vacuum oven.
15. Resuspend the DNA in 10:0.1 TE buffer [11].

Maxam–Gilbert sequencing

In 1976–1977, Allan Maxam and Walter Gilbert developed a DNA sequencing method based on chemical modification of DNA and subsequent cleavage at specific bases. Although Maxam and Gilbert published their chemical sequencing method two years after the ground-breaking paper of Sanger and Coulson on plus-minus sequencing, Maxam–Gilbert sequencing rapidly became more popular, since purified DNA could be used directly, while the initial Sanger method required that each read start be cloned for production of single-stranded DNA. However, with the improvement of the chain-termination method (see below), Maxam-Gilbert sequencing has fallen out of favour due to its technical complexity prohibiting its use in standard molecular biology kits, extensive use of hazardous chemicals, and difficulties with scale-up [12].

The method requires radioactive labeling at one 5' end of the DNA (typically by a kinase reaction using gamma-^{32}P ATP) and purification of the DNA fragment to be sequenced. Chemical treatment generates breaks at a small proportion of one or two of the four nucleotide bases in each of four reactions (G, A+G, C, C+T). For example, the purines (A+G) are depurinated using formic acid, the guanines (and to some extent the adenines) are methylated by dimethyl sulfate, and the pyrimidines (C+T) are methylated using hydrazine. The addition of salt (sodium chloride) to the hydrazine reaction inhibits the methylation of thymine for the C-only reaction. The modified DNAs are then cleaved by hot piperidine at the position of the modified base. The concentration of the modifying chemicals is controlled to introduce on average one modification per DNA molecule. Thus a series of labeled fragments is generated, from the radiolabeled end to the first "cut" site in each molecule. The fragments in the four reactions are electrophoresed side by side in denaturing acrylamide gels for size separation. To visualize the fragments, the gel is exposed to X-ray film for autoradiography, yielding a series of dark bands each corresponding to a radiolabeled DNA fragment, from which the sequence may be inferred. Also sometimes known as "chemical sequencing", this method led to the Methylation Interference Assay used to map DNA-binding sites for DNA-binding proteins.

Part of a radioactively labeled sequencing gel

Because the chain-terminator method (or Sanger method after its developer Frederick Sanger) is more efficient and uses fewer toxic chemicals and lower amounts of radioactivity than the method of Maxam and Gilbert, it rapidly became the method of choice. The key principle of the Sanger method was the use of dideoxynucleotide triphosphates (ddNTPs) as DNA chain terminators.

The classical chain-termination method requires a single-stranded DNA template, a DNA primer, a DNA polymerase, normal deoxynucleotide triphosphates (dNTPs), and modified nucleotides (dideoxy NTPs) that terminate DNA strand elongation. These ddNTPs will also be radioactively or fluorescently labeled for detection in automated sequencing machines. The DNA sample is divided into four separate sequencing reactions, containing all four of the standard deoxy nucleotides (dATP, dGTP, dCTP and dTTP) and the DNA polymerase. To each reaction is added only one of the four dideoxy nucleotides (ddATP, ddGTP, ddCTP, or ddTTP) which are the chain-terminating nucleotides, lacking a 3'-OH group required for the formation of a phosphodiester bond between two nucleotides, thus terminating DNA strand extension and resulting in DNA fragments of varying length [13].

The newly synthesized and labelled DNA fragments are heat denatured, and separated by size (with a resolution of just one nucleotide) by gel electrophoresis on denaturing polyacrylamide-urea gel with each of the four reactions run in one of four individual lanes (lanes A, T, G, C); the DNA bands are then visualized by autoradiography or UV light, and the DNA sequence can be directly read off the X-ray film or gel image. In the image on the right, X-ray film was exposed to the gel, and the dark bands correspond to DNA fragments of different lengths. A dark band in a lane indicates a DNA fragment that is the result of chain termination after incorporation of a dideoxynucleotide (ddATP, ddGTP, ddCTP, or ddTTP). The relative positions of the different bands among the four lanes are then used to read (from bottom to top) the DNA sequence.

DNA fragments are labeled with a radioactive or fluorescent tag on the primer, in the new DNA strand with a labeled dNTP, or with a labeled ddNTP. Technical variations of chain-termination sequencing include tagging with nucleotides containing radioactive phosphorus for radio labeling, or using a primer labeled at the 5' end with a fluorescent dye. Dye-primer sequencing facilitates reading in an optical system for faster and more economical analysis and automation. The later development by Leroy Hood and coworkers of fluorescently labeled ddNTPs and primers set the stage for automated, high-throughput DNA sequencing [14].

Chain-termination methods have greatly simplified DNA sequencing. For example, chain-termination-based kits are commercially available that contain the reagents needed for sequencing, pre-aliquoted and ready to use. Limitations include non-specific binding of the primer to the DNA, affecting accurate read-out of the DNA sequence, and DNA secondary structures affecting the fidelity of the sequence [15].

Dye-terminator sequencing

Dye-terminator sequencing utilizes labeling of the chain terminator ddNTPs, which permits sequencing in a single reaction, rather than four reactions as in the labelled-primer method. In dye-terminator sequencing, each of the four dideoxy nucleotide chain terminators is labeled with fluorescent dyes, each of which emit light at different wavelengths. Owing to its greater expediency and speed, dye-terminator sequencing is now the mainstay in automated sequencing. Its limitations include dye effects due to differences in the incorporation of the dye-labeled chain terminators into the DNA fragment, resulting in unequal peak heights and shapes in the electronic DNA sequence trace chromatogram after capillary electro phoresis.

This problem has been addressed with the use of modified DNA polymerase enzyme systems and dyes that minimize incorporation variability, as well as methods for eliminating "dye blobs". The dye-terminator sequencing method, along with automated high-throughput DNA sequence analyzers, is now being used for the vast majority of sequencing projects.

•The human genome's gene-dense "urban centers" are in nucleotides G , A and T ,C.
•In contrast, the gene-poor "deserts" are rich in the DNA nucleotides.

DNA ASSEMBLING

Ligase mechanism

The mechanism of DNA ligase is to form two covalent phospho diester bonds between 3' hydroxyl ends of one nucleotide, ("acceptor") with the 5' phosphate end of another ("donor"). ATP is required for the ligase reaction, which proceeds in three steps: (1) adenylation (addition of AMP) of a residue in the active center of the enzyme, pyrophosphate is released; (2) transfer of the AMP to the 5' phosphate of the so-called donor, formation of a pyrophosphate bond; (3) formation of a phosphodiester bond between the 5' phosphate of the donor and the 3' hydroxyl of the acceptor. Ligase will also work with blunt ends, although higher enzyme concentrations and different reaction conditions are required.

Mammalian ligases

- In mammals, there are four specific types of ligase.
- DNA ligase I: ligates the nascent DNA of the lagging strand after the Ribonuclease H has removed the RNA primer from the Okazaki fragments.
- DNA ligase II: alternatively spliced form of DNA ligase III found in non-dividing cells.
- DNA ligase III: complexes with DNA repair protein XRCC1 to aid in sealing DNA during the process of nucleotide excision repair and recombinant fragments.
- DNA ligase IV: complexes with XRCC4. It catalyzes the final step in the non-homologous end joining DNA double-strand break repair pathway. It is also required for V(D)J recombination, the process that generates diversity in immunoglobulin and T-cell receptor loci during immune system development.

Some forms of DNA ligase present in bacteria (usually larger) may require NAD to act as a co-factor, whereas other forms of DNA ligases (usually present in *E.coli*, and usually smaller) may require ATP to react. Also, a number of other structures present in the DNA ligase are the AMP and lysine, both of which are important in the ligation process since they create an intermediate enzyme.

Draft human genome sequence

- Genes appear to be concentrated in random areas along the genome, with vast expanses of non-codingDNA between Chromosome 1 has the most genes (2968), and the Y chromosome has the fewest (231)
- Less than 2% of the genome codes for proteins.
- Repeated sequences that do not code for proteins ("junk DNA") make up at least 50% of the human genome. Repetitive sequences shed light on chromosome structure and dynamics. Over time, these repeats reshape the genome by rearranging it, creating entirely new genes, and modifying and reshuffling existing genes [16].

The repeats fall into five classes:

- Transposon-derived repeats, known as interspersed repeats.
- Inactive retroposed copies of cellular genes, known as processed pseudogenes. Nonfunctional copies of the exon sequences of an active gene and thought to arise by integration into chromosomes of a natural cDNA sequence generated by reverse transcription.
- Repeats of short k-mers such as (A)n, (CA)n, (AAT)n. Since they show a high degree of length polymorphisms in the human population, (CA)n repeat have been used as genetic marker in genetic mapping.
- Segmental duplications, consisting of blocks of 10-300 kb that have been copied from one region of the genome into another region. Such duplications appears often in pericentromeres and sub telomeres of chromosomes. Recurrent structural rearrangements in duplication regions give rise to contiguous gene syndromes.
- Tandemly repeated sequences, usually at centromere, telomers, the short arms of acrocentric chromosomes and ribosomal gene clusters. These regions are under-represented in the draft genome sequence.

SUB CLONING :

- DNA sequencers can only read small fragments of DNA 500-1000 bases longIt is necessary to break the genome into small pieces.
- Individual chromosomes are cut into ~1 million base chunks that are cloned into large vectors called BACs, PACs, and YACs.
- These pieces can then be further cut into sequence able pieces (~1000 bases) and cloned into plasmid or phage vectors.

THE BENEFITS OF HUMAN GENOME PROJECT

With all these consequences to the project, one has to wonder if there are any good results due to the research. So let’s look at the benefits of the Human Genome Project. Currently, there are 500 genetic tests as a result of the project. As of now, they are being used in laboratories in order to study families with known genetic diseases. Results can show either predisposition to a disease or to analyze differences between healthy and ill patients. Only 50 of these 500 tests are performed in clinics until more information is gathered.

Additional benefits include identifying genes for specific diseases, such as Parkinson’s disease, breast cancer, polycystic kidney diseases and types of deafness. Researchers have done this by finding misspellings in the sequence of DNA and then having biotechnology companies develop tests. Also, there has been improvement in drug design and organ replacement.

Scientists have come a long way since the beginning. For example, it now costs less than 10 cents to identify a bit of DNA, yet two years ago it cost 2 dollars. Imagine how much that saves when humans have billions of pieces of DNA. Also, the technology to perform these tests has become much more sophisticated. Laboratories now have DNA probe tests that automatically find the DNA of an organism that causes a disease. Chlamydia and tuberculosis were just two of the diseases found this way.

The Human Genome Project is a great tool for medicine, but more knowledge and acceptability will be the greatest tool. People have to realize that there are two sides to every issue. Benefits and costs must be weighed against one another, limits must be set, and ethical standards met. Already, the government has intervened and provided some guidelines. Among many other laws passed, The Health Insurance Portability and Accountability Act was passed in 1996. A major part of the law was to stop the denial of health coverage on the basis of genetic information.

The work on interpretation of genome data is still in its initial stages. It is anticipated that detailed knowledge of the human genome will provide new avenues for advances in medicine and biotechnology.

Clear practical results of the project emerged even before the work was finished. For example, a number of companies, such as Myriad Genetics started offering easy ways to administer genetic tests that can show predisposition to a variety of illnesses, including breast cancer, hemostasis disorders, cystic fibrosis, liver diseases and many others. Also, the etiologies for cancers, Alzheimer's disease and other areas of clinical interest are considered likely to benefit from genome information and possibly may lead in the long term to significant advances in their management.

There are also many tangible benefits for biological scientists. For example, a researcher investigating a certain form of cancer may have narrowed down his/her search to a particular gene. By visiting the human genome database on the World Wide Web, this researcher can examine what other scientists have written about this gene, including (potentially) the three-dimensional structure of its product, its function(s), its evolutionary relationships to other human genes, or to genes in mice or yeast or fruit flies, possible detrimental mutations, interactions with other genes, body tissues in which this gene is activated, diseases associated with this gene or other datatypes.

Further, deeper understanding of the disease processes at the level of molecular biology may determine new therapeutic procedures. Given the established importance of DNA in molecular biology and its central role in determining the fundamental operation of cellular processes, it is likely that expanded knowledge in this area will facilitate medical advances in numerous areas of clinical interest that may not have been possible without them.

The analysis of similarities between DNA sequences from different organisms is also opening new avenues in the study of evolution. In many cases, evolutionary questions can now be framed in terms of molecular biology; indeed, many major evolutionary milestones (the emergence of the ribosome and organelles, the development of embryos with body plans, the vertebrate immune system) can be related to the molecular level. Many questions about the similarities and differences between humans and our closest relatives (the primates, and indeed the other mammals) are expected to be illuminated by the data from this project.

The Human Genome Diversity Project (HGDP), spinoff research aimed at mapping the DNA that varies between human ethnic groups, which was rumored to have been halted, actually did continue and to date has yielded new conclusions.[*citation needed*] In the future, HGDP could possibly expose new data in disease surveillance, human development and anthropology. HGDP could unlock secrets behind and create new strategies for managing the vulnerability of ethnic groups to certain diseases (see race in biomedicine). It could also show how human populations have adapted to these vulnerabilities [17].

1. *Medical Benefits*

- Improved diagnosis of disease eg cancer, high blood pressure
- Detection of genetic predisposition to disease
- Drug design and gene therapy
- When faulty genes are detected the disease can be treated early, it also leads to diagnostic tests which can include genetic counseling
- Normal genes may be cloned and the product of the expression of these genes used for treatment.

2. *Non-Medical Benefits*

- Creates understanding of human evolution eg comparison between species, understanding of evolutionary relationships (eg 1% difference between us and chimps)
- Development in forensics

Advantages of Human Genome Project:

1. Knowledge of the effects of variation of DNA among individuals can revolutionize the ways to diagnose, treat and even prevent a number of diseases that affects the human beings.
2. It provides clues to the understanding of human biology.

APLLICATIONS

Public Health Applications

In virtually all of the reviews, it was concluded that there was no clear immediate public health application of the data. However, several of the reviews highlighted gene-disease associations for which public health applications are being considered. For example, one review dealt with sickle cell disease, for which an intervention has been established following a randomized trial that showed that oral penicillin could significantly reduce the associated morbidity and mortality. The substantial differences in mortality due to sickle cell disease that were demonstrated may reflect differences in the timing of introduction and extent of coverage of newborn screening and differences in medical care, parental education, and penicillin prophylaxis to prevent infections. Another review considered *HLA-DQ* and type 1 diabetes, as well as the weight of evidence that has led to *HLA-DQ* screening for type 1 diabetes being conducted in high-risk families and the general population for intervention trials and natural history studies. The review also highlighted a critical need to reconsider the risks, benefits, and ethical, legal, and social issues regarding genetic or autoantibody testing for type 1 diabetes, as well as a need to clarify the effects of environmental exposures as independent or interacting with high-risk *HLA* genotypes. In the HuGE review of hereditary hemachromatosis, it was concluded that more information is needed about penetrance of clinical expression among persons with elevated transferrin saturation or *HFE* mutations, about the disease burden associated with hereditary hemachromatosis in the general population, about screening accuracy, and about the diagnostic tests available and the efficacy of early treatment. For medium chain acyl coenzyme A dehydrogenase deficiency, the main knowledge gap concerns the natural history of the disease and its clinical outcomes. In regard to mismatch repair genes and colorectal cancer, there is no consensus regarding the mostefficient approach of identifying mutation carriers. Some of the other reviews dealt with gene variants that are part of a number of biomarkers included in test kits being marketed commercially, and these reviews highlighted important gaps in the evidence base. In future reviews, we encourage authors to emphasize data gaps and make recommendations for research to address these gaps [18].

VARIATION IN MANIFESTATION

In all of the reviews dealing with gene variants associated with a high risk of disease, variable penetrance or manifestation was noted. This reinforces the point that even for single gene disorders there is wide variation in clinical phenotype (18), and for this reason HuGE reviews of these disorders are valuable. In all of the reviews, there was a lack of data on other factors contributing to variation in manifestation.

METHODOLOGICAL ISSUES

The reviews highlight methodological issues such as selection bias, statistical power, and investigation of interaction or modifying factors, and they uncovered a need for unified guidelines that can be used to synthesize results of the increasing number of such studies. Progress is being made in defining quality standards for genetic-epidemiologic research, but ongoing evaluation is needed to make sure that such guidelines are refined and implemented. In 2001, an expert panel sponsored by the Centers for Disease Control and Prevention and the National Institutes of Health developed guidelines and recommendations for the reporting, evaluation, and integration of data from human genome epidemiology with emphasis on studies of 1) prevalence of gene variants and gene-disease associations, 2) gene-environment and gene-gene interactions, and 3) evaluation of genetic tests. Conclusions and recommendations from this workshop have been published. In addition, other groups have proposed guidelines for gene-disease association studies. Many of the recommendations are similar, and the use of these guidelines in reporting studies should facilitate the integration of evidence in the future. Similarly, there is increasing interest in standardized approaches to the evaluation of genetic tests [14].

QUANTITATIVE SYNTHESIS

The use of meta-analysis or pooled analysis as a tool to synthesize evidence has been left to the discretion of the authors of HuGE reviews, in part because of concern about the lack of comparability of study methods and in part because of concern about the validity of meta-analysis of observational studies. Meta-analysis was used as a tool for synthesizing evidence in two of the reviews. In the future, with the application of guidelines for reporting human genome epidemiology studies, more comparability among published data will make meta-analysis a more feasible option. For the present, as the potential value of using meta-analysis is likely to vary between different gene-disease associations, we prefer to leave this decision to the authors of reviews. In one of the reviews, pooled analysis (which requires data on individual subjects) was used in addition to meta-analysis. Interestingly, the results of the pooled and meta-analyses were very similar. Pooled analyses require much greater resources than meta-analyses and would be preferred to meta-analysis only when a high degree of precision of the measures of effect is required. For example, as data on the entrance of HFE mutations accumulate, a pooled analysis might be of considerable value.

REPLICATION

More generally, there has been considerable concern about nonreplication of gene-disease association studies. Nonreplication has also been an issue in other areas of epidemiologic research, so much so that epidemiology has been occasionally viewed as having reached its limits ; for example, the results of recent cohort studies are challenging the inverse association between cancer and consumption of vegetables and fruit. The investigation of gene-disease associations differs from the investigation of exposure-disease associations in two important respects. First, the assessment of genotypes by DNA assays (polymerase chain reaction methods) is generally more accurate than for exposure assessment, and it is less heavily dependent on study design. Second, because of “Mendelian randomization”, an association between a disease and a genotype is unlikely to be due to confounding, provided that the study is designed according to the principles of population-based studies. Although there has been concern about population stratification, empirical studies in non-Hispanic White Americans and modeling suggest that bias from this source may not be substantial when epidemiologic principles of study design, conduct, and analysis are rigorously applied. In this context, it is interesting that, in an analysis of 301 published studies covering 25 associations in which the first positive report was excluded, grouping studies by ethnicity generally did not remove heterogeneity. In the same meta-analysis, there was an excess of studies replicating the initial report that seemed unlikely to be due to publication bias. For eight of the associations, the combined estimate of relative risk was statistically significant; this proportion is similar to the findings of another set of meta-analyses. Thus, it is possible that, as an area of investigation matures with a move from small innovative studies which might best be viewed as pilot studies to large well-resourced studies in which potential biases are minimized, more consistent associations will be observed than predicted by the rather bleak commentaries based on early studies.

This raises the challenge of keeping overviews of evidence up-to-date. In the early stages of an area of investigation, publication bias may be of critical importance, as suggested for example by the pattern of accrual of evidence regarding the association between the angiotensin – converting - enzyme insertion/deletion polymorphism and myocardial infarction. Differences in timing may account for some differences between the results of meta-analyses as evidence accrues. Later, publication bias may be less of an issue as large high-quality studies are likely to be published irrespective of their findings. The best solution to the problem of publication bias appears to be the establishment of a research register for studies of gene-disease associations and of gene-gene and gene-environment interactions, analogous to those for other areas of medicine. This would help to address the problem of integrating all available evidence, taking into account its quality [16].

VOLUME AND TYPE OF EVIDENCE

There is also the challenge of the ever-increasing number of human genome epidemiology studies. For example, in the literature database maintained in the Centers for Disease Control and Prevention Genomics and Disease Prevention Information System, 2,436 primary studies of this type were published in 2001, and 2,922 studies were published in 2002. Moreover, as a result of the increasing availability of mapped single nucleotide polymorphism markers, this trend is expected to accelerate. Therefore, integration of evidence will become increasingly important as a means of dealing with potentially unmanageable amounts of information. Certainly, the Human Genome Epidemiology Network (HuGE Net) will continue to benefit from the contributions of researchers writing HuGE reviews in their own specialty areas. However, we would also like to suggest some priorities with the hope of encouraging others to invest effort in integrating evidence about the gene-disease associations (and related gene-gene and gene-environment interactions) most likely to expand our knowledge and ability to apply research results.

Past discovery

- Just a half-century ago, very little was known about the genetic factors that contribute to human disease.
- In 1953, James Watson and Francis Crick described the double helix structure of deoxyribonucleic acid (DNA), the chemical compound that contains the genetic instructions for building, running and maintaining living organisms.
- Methods to determine the order, or sequence, of the chemical letters in DNA were developed in the mid-1970s.
- In 1990, the National Institutes of Health (NIH) and the Department of Energy joined with international partners in a quest to sequence all 3 billion letters, or base pairs, in the human genome, which is the complete set of DNA in the human body. This concerted, public effort was the Human Genome Project.
- The Human Genome Project’s goal was to provide researchers with powerful tools to understand the genetic factors in human disease, paving the way for new strategies for their diagnosis, treatment and prevention.
- From the start, the Human Genome Project supported an Ethical, Legal and Social Implications research program to address the many complex issues that might arise from this science.
- All data generated by the Human Genome Project were made freely and rapidly available on the Internet, serving to accelerate the pace of medical discovery around the globe.
- The Human Genome project spurred a revolution in biotechnology innovation around the world and played a key role in making the U.S. the global leader in the new biotechnology sector.
- In April 2003, researchers successfully completed the Human Genome Project, under budget and more than two years ahead of schedule [17].

Present discovery

- The Human Genome Project has already fueled the discovery of more than 1,800 disease genes.
- As a result of the Human Genome Project, today’s researchers can find a gene suspected of causing an inherited disease in a matter of days, rather than the years it took before the genome sequence was in hand.
- There are now more than 2,000 genetic tests for human conditions. These tests enable patients to learn their genetic risks for disease and also help healthcare professionals to diagnose disease.
- At least 350 biotechnology-based products resulting from the Human Genome Project are currently in clinical trials.
- Having the complete sequence of the human genome is similar to having all the pages of a manual needed to make the human body. The challenge now is to determine how to read the contents of these pages and understand how all of these many, complex parts work together in human health and disease.
- One major step toward such comprehensive understanding was the development in 2005 of the HapMap (http://hapmap.ncbi.nlm.nih.gov/), which is a catalog of common genetic variation, or haplotypes, in the human genome. In 2010, the third phase of the HapMap project was published, with data from 11 global populations, the largest survey of human genetic variation performed to date. HapMap data have accelerated the search for genes involved in common human diseases, and have already yielded impressive results in finding genetic factors involved in conditions ranging from age-related blindness to obesity.
- The tools created through the Human Genome Project continue to underlie efforts to characterize the genomes of important organisms used extensively in biomedical research, including fruit flies, roundworms, and mice.
- NIH’s Ethical, Legal and Social Implications program has become a model for other research efforts seeking to address ethical issues in a proactive manner (http://www.genome.gov/10001618).
- With the drastic decline in the cost of sequencing whole exomes or genomes, groundbreaking comparative genomic studies are now identifiying the causes of rare diseases such as Kabuki and Miller syndromes.

- Much work still remains to be done. Despite many important genetic discoveries, the genetics of complex diseases such as heart disease are still far from clear.
- Pharmacogenomics is a field that looks at how genetic variation affects an individual’s response to a drug. Pharmacogenomic tests can already identify whether or not a breast cancer patient will respond to the drug Herceptin, whether an AIDS patient should take the drug Abacavir, or what the correct dose of the blood-thinner Warfarin should be.

Future discovery

- An ambitious new initiative, The Cancer Genome Atlas (http://cancergenome.nih.gov/), aims to identify all the genetic abnormalities seen in 50 major types of cancer.
- Based on a deeper understanding of disease at the genomic level, we will see a whole new generation of targeted interventions, many of which will be drugs that are much more effective and cause fewer side effects than those available today.
- NIH-supported access to high-throughput screening of small molecule libraries will provide academic researchers with powerful new research probes to explore the hundreds of thousands of proteins believed to be encoded by the approximately 25,000 genes in the human genome, and will provide innovative techniques to spur development of new, more effective, types of drugs.
- NIH is striving to cut the cost of sequencing an individual’s genome to $1,000 or less. Having one’s complete genome sequence will make it easier to diagnose, manage and treat many diseases.
- Individualized analysis based on each person’s genome will lead to a powerful form of preventive, personalized and preemptive medicine. By tailoring recommendations to each person’s DNA, health care professionals will be able to work with individuals to focus efforts on the specific strategies — from diet to high-tech medical surveillance — that are most likely to maintain health for that particular individual.
- The increasing ability to connect DNA variation with non-medical conditions, such as intelligence and personality traits, will challenge society, making the role of ethical, legal and social implications research more important than ever [15,18].

Fig 1. Structure of DNA

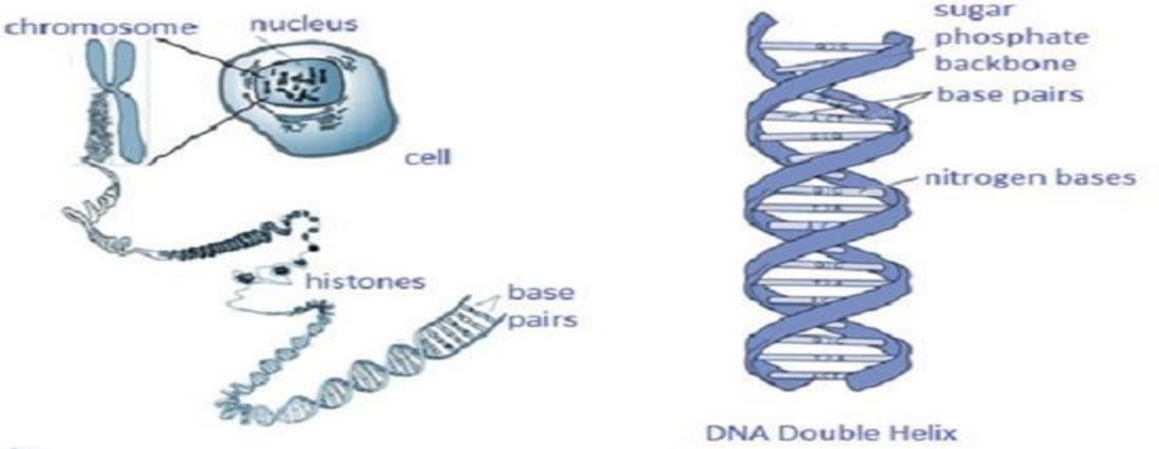
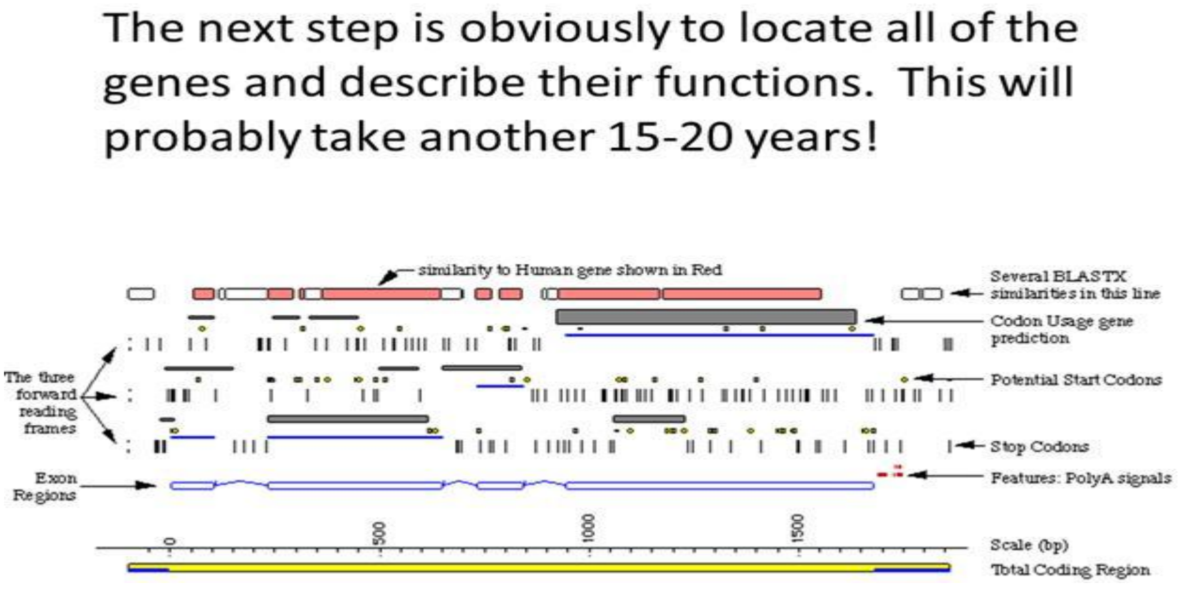


Fig 2. Genome sequence


The next step is obviously to locate all of the genes and describe their functions. This will probably take another 15-20 years!

similarity to Human gene shown in Red
Several BLASTX similarities in this line
Codon Usage gene prediction
The three forward reading frames
Potential Start Codons
Stop Codons
Exon Regions
Features: PolyA signals
0
500
1000
1500
Scale (bp)
Total Coding Region

Fig 3. Hierarchical shotgun sequencing

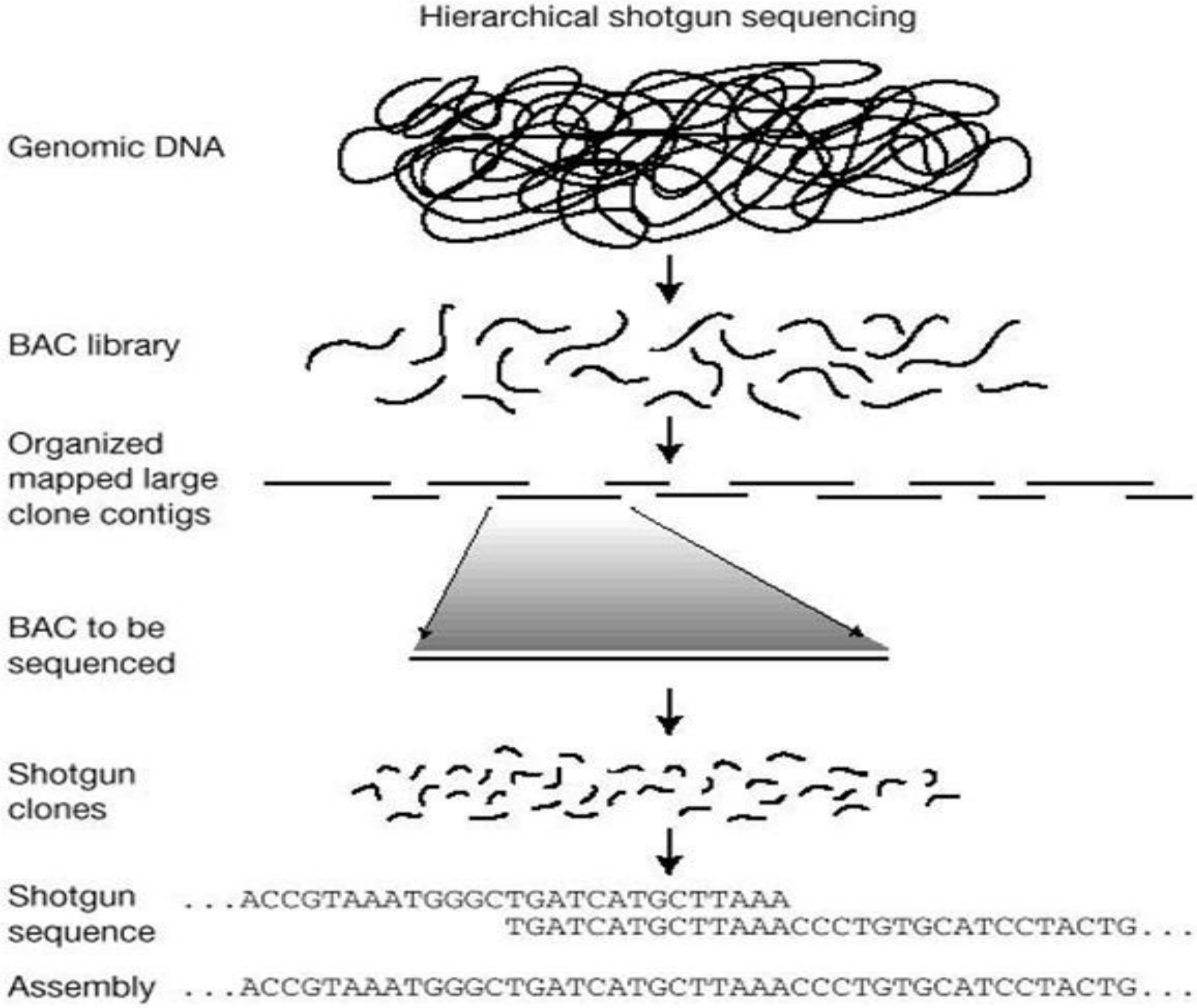
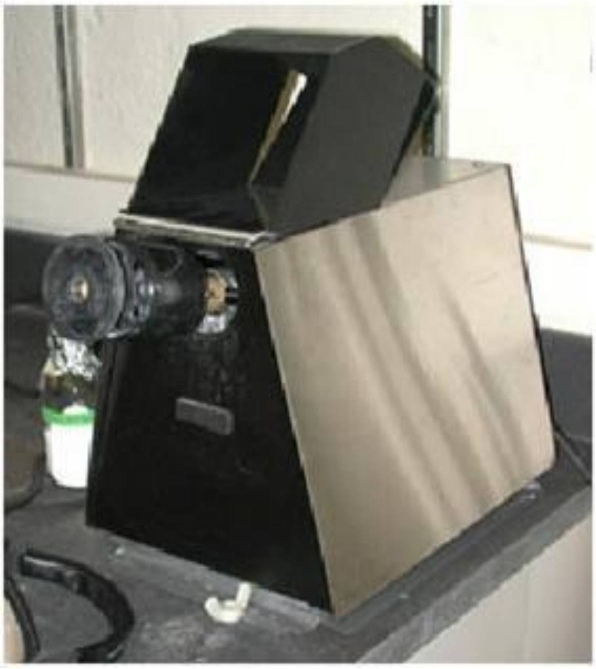
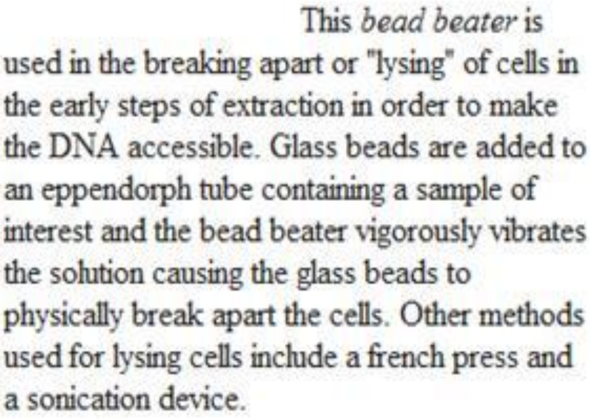

Hierarchical shotgun sequencing
Genomic DNA
BAC library
Organized mapped large clone contigs
BAC to be sequenced
Shotgun clones
Shotgun sequence
...ACCGTAAATGGGCTGATCATGCTTAAA
TGATCATGCTTAAACCCTGTGCATCCTACTG...
Assembly
...ACCGTAAATGGGCTGATCATGCTTAAACCCTGTGCATCCTACTG...

Fig 4. Bead beater





This *bead beater* is used in the breaking apart or "lysing" of cells in the early steps of extraction in order to make the DNA accessible. Glass beads are added to an eppendorph tube containing a sample of interest and the bead beater vigorously vibrates the solution causing the glass beads to physically break apart the cells. Other methods used for lysing cells include a french press and a sonication device.

Fig 5. Centrifuger

A *centrifuge* such as this can spin at up to 15,000 rpm to facilitate separation of the different phases of the extraction. It is also used to precipitate the DNA after the salts are washed away with ethanol and or isopropanol.

Fig 6. Gel box

A ***gel box*** is used to separate DNA in an the net negative charge of the molecule. Different sized pieces of DNA move at different rates, with the larger pieces moving more slowly through the porus medium, thereby creating a size separation that can be differentiated in a gel.

Fig 7. Decoding genes

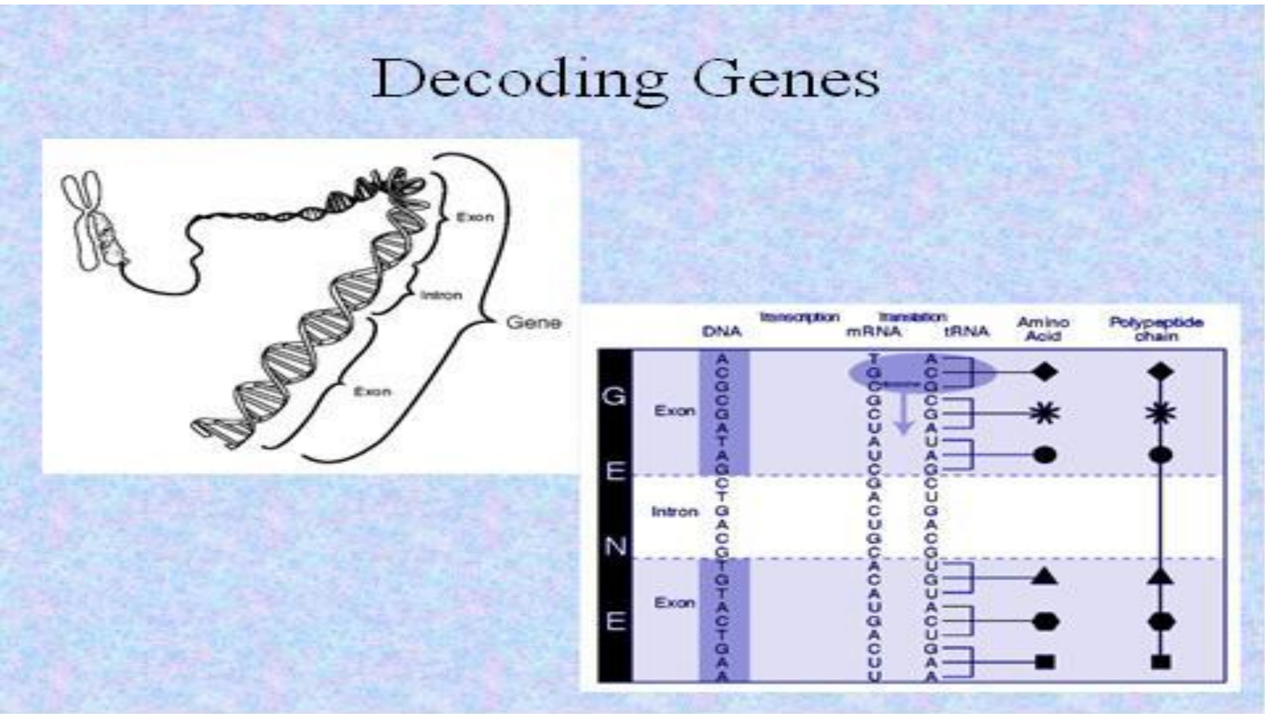
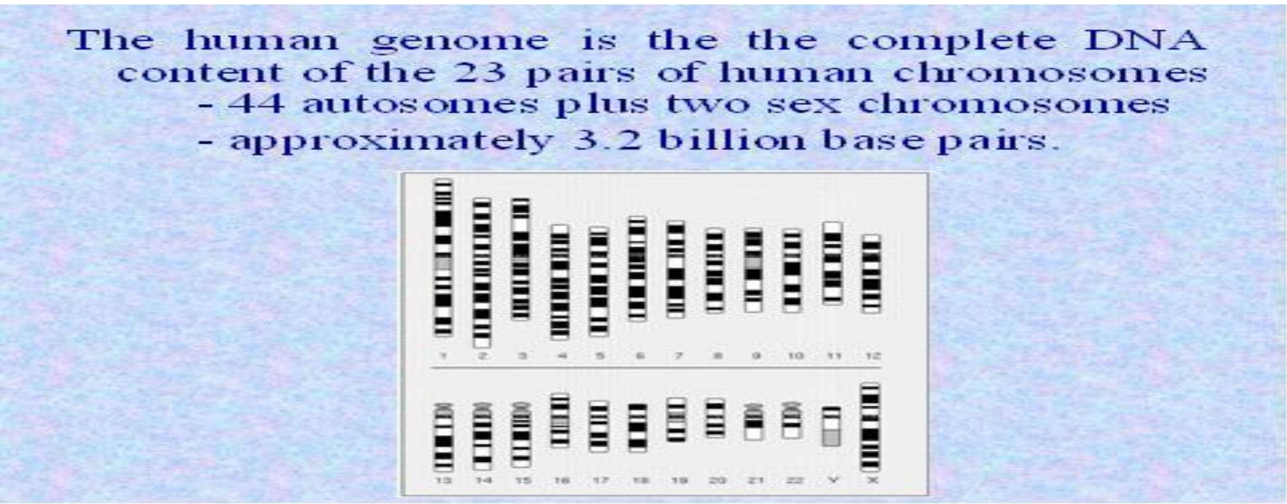


Fig 8. The human genome



The human genome is the the complete DNA content of the 23 pairs of human chromosomes
- 44 autosomes plus two sex chromosomes
- approximately 3.2 billion base pairs.

Fig 9. DNA – Structure

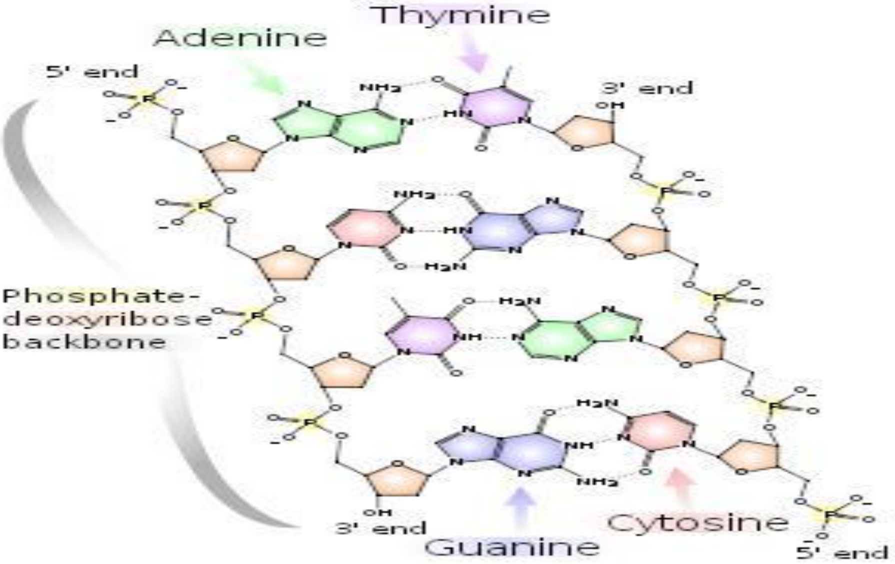


Fig 10. Maxam–Gilbert sequencing

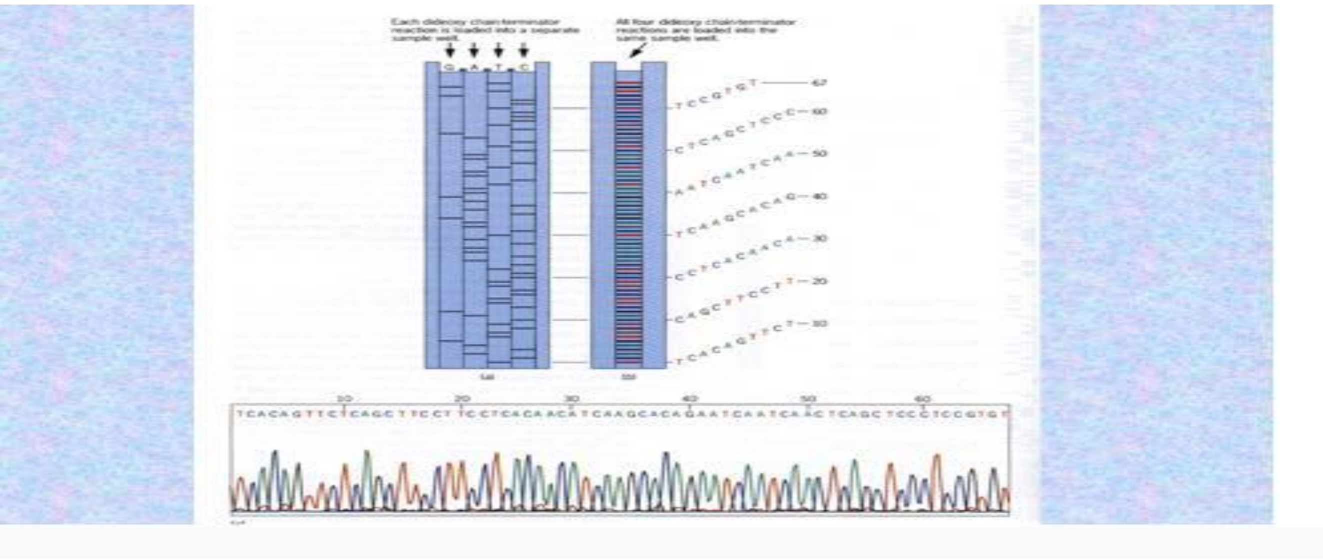

Each dideoxy chain-terminator reaction is loaded into a separate sample well.
All four dideoxy chain-terminator reactions are loaded into the same sample well.
G
A
T
C
67
60
50
40
30
20
10
(a)
(b)
10
20
30
40
50
60

Fig 11. Radioactive Labeled Sequencing Gel

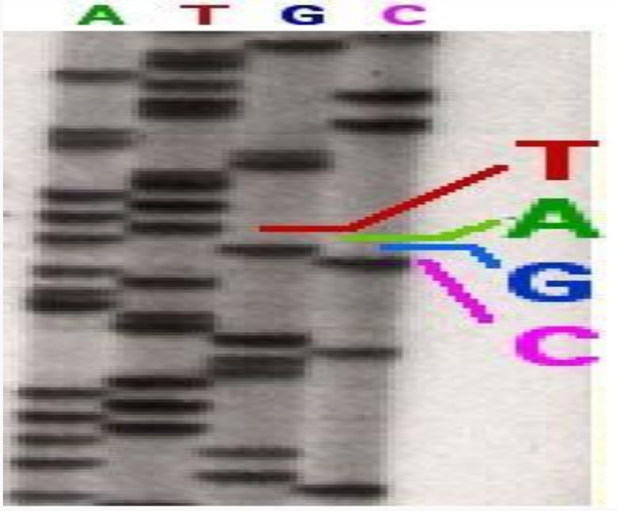


Fig 12. Labeled DNA sequence

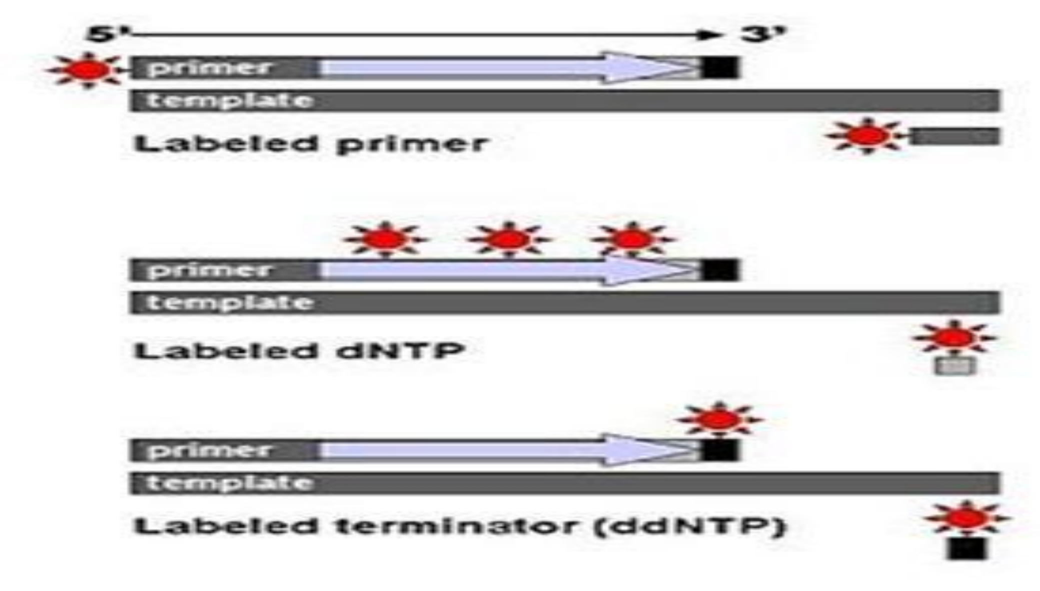


Fig 13. Sequence ladder by radioactive sequencing compared to fluorescent peaks

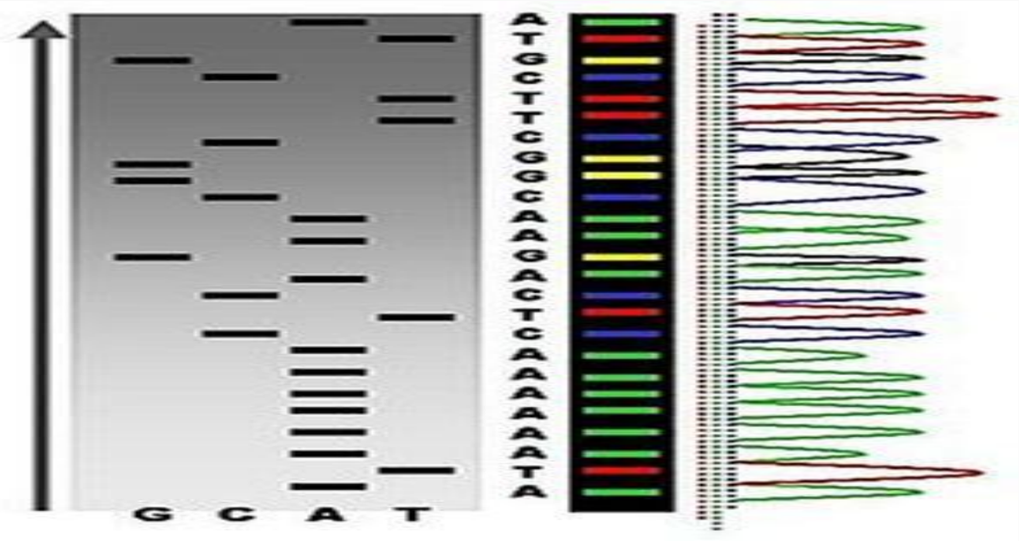
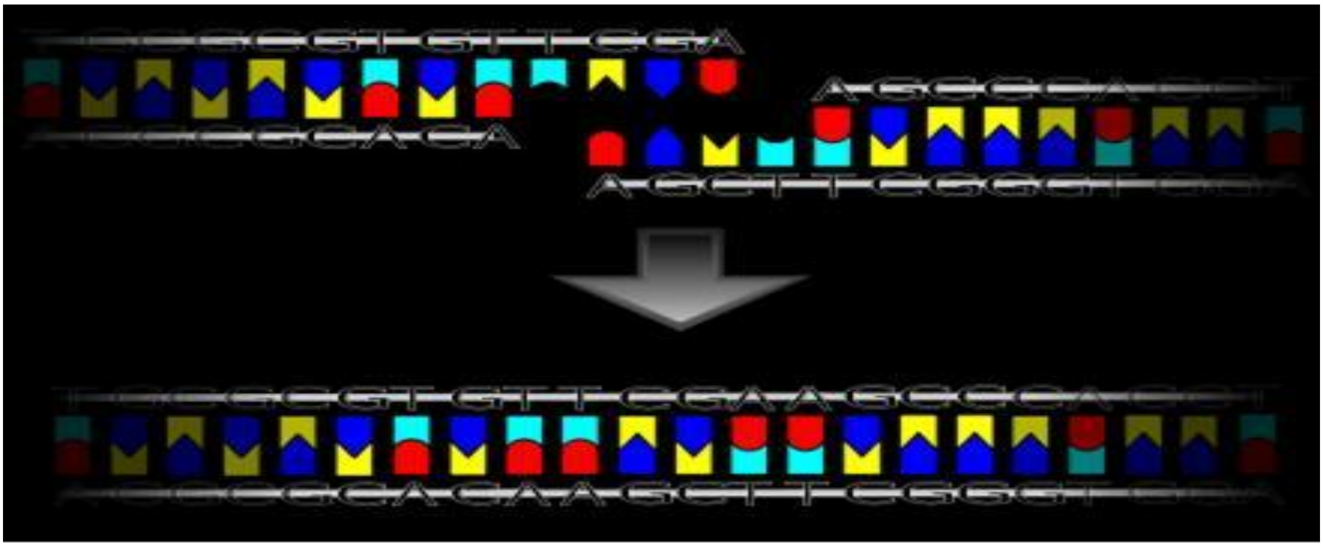


Fig 14. DNA ligase



CONCLUSION

Analysis of the draft sequence revealed a vast amount of information including: The average human gene consists of 3,000 nucleotide bases, but sizes vary greatly – the largest known human gene has 2.4 million bases, The order of 99.9% of nucleotide bases is exactly the same in all people, The functions of over 50% of discovered genes remain unknown, Less than 2% of the genome encodes for the production of proteins, Much of the genome consists of repetitive base sequences. These repeats appear to have no direct function, but over time reshape the genome by rearranging it; creating new genes or modifying and reshuffling existing genes, Gene-rich areas of the genome are predominantly made up of G and C bases, whereas gene-poor regions are mainly composed of A and T bases, Chromosome 1 has the most genes (2968) whereas the Y chromosome has the least (231). Much is still unknown about our genome. Some of the things we still don't know are: The exact number of genes in the human genome, The exact location, function and regulation of these genes, The amount, distribution, information content and functions of 'non-coding' DNA, that is, DNA that does not code for a protein product, How gene expression, protein expression and post-translational events are orchestrated, Evolutionary conservation of genes and proteins amongst different organisms, Correlation of genetic variation between individuals with respect to health and disease.

REFERENCES

1. Robert Krulwich. *Cracking the Code of Life*. About the Human Genome Project: What is the Human Genome Project. The HuMan Genome Management Information System (HGMIS). 2011-07-18. Retrieved 2011-09-02.
2. Harmon, Katherine. Genome Sequencing for the Rest of Us. Scientific American. Retrieved 2010-08-13.
3. Cook-Deegan R. The Alta Summit, December 1984". *Genomics,* 5 (3), 1989, 661–3.
4. Barnhart, Benjamin J. DOE Human Genome Program. *Human Genome Quarterly*, 1, 1989, 1.
5. DeLisi, Charles. Genomes: 15 Years Later A Perspective by Charles DeLisi, HGP Pioneer. *Human Genome News,* 11, 2001, 3–4.
6. Noble, Ivan. Human genome finally complete. *BBC News*. Retrieved 2006-07-22.
7. Anonumous 1. http://www.strategicgenomics.com/Genome/index.htm
8. Anonumous 2. http://genome.ucsc.edu
9. Anonumous 3. http://www.ensembl.org
10. Roach Boysen, C, Wang K, Hood L. Pairwise end sequencing: a unified approach to genomic mapping and sequencing. *Genomics,* 26 (2), 1995, 345–353.
11. The Human Genome Project Race. Center for Biomolecular Science and Engineering. Retrieved 2011-05-02.
12. Venter, JC et al. The sequence of the human genome. *Science,* 291 (5507), 2001, 1304–1351.
13. IHGSC. Finishing the euchromatic sequence of the human genome. *Nature* 431 (7011), 2004, 931–945.
14. Osoegawa Kazutoyo, Mammoser AG, Wu C, Frengen E, Zeng C, Catanese JJ, De Jong PJ. A Bacterial Artificial Chromosome Library for Sequencing the Complete Human Genome. *Genome Research,* 11 (3), 2001, 483–96.
15. Tuzun, E et al. Fine-scale structural variation of the human genome. *Nature Genetics,* 37 (7), 2005, 727–737.
16. Kennedy D. Not wicked, perhaps, but tacky. *Science,* 297 (5585), 2002, 1237.
17. Venter D. A Part of the Human Genome Sequence. *Science,* 299 (5610), 2003, 1183–4.
18. Levy S, Sutton G, Ng PC et al. The Diploid Genome Sequence of an Individual Human. *PLoS Biol.,* 5 (10), 2007, e254.