The Human Genome Project

YOSSI SEGAL 1996

THE ISRAEL ACADEMY OF SCIENCES AND HUMANITIES

FOREWORD

The concept of deciphering the Human Genome surfaced in the United States in the 1930s following the discoveries that color blindness and hemophilia are linked to chromosome X.

The Human Genome Project (HGP) originated in the U.S. at the Department of Energy (DOE) meeting in Alta, Utah in December 1984, when the possible use of DNA analysis in detecting mutations among atomic bomb survivors was contemplated.

Following lengthy deliberations, the U.S. government approved the program, and in 1988 the HGP was launched under the supervision of the National Institutes of Health (NIH) and DOE. In 1990 it was shaped into the form of a 15-year program, designed to map and sequence the entire Human Genome and also of several model organisms, at a yearly budget of two hundred million dollars totaling three billion dollars, to end in the year 2005.

Several other countries, headed by France, the UK and to a lesser extent Japan, have joined this effort. Contiguous to this is an on going international collaboration of many other countries, including Israel. It also includes the contribution of several international organizations, such as HUGO (Human Genome Organization), the European Community (EC), and UNESCO, in the dissemination of knowledge and information and the support of scientific and technological activities.

The present publication provides a concise description of the evolvement of the HGP, its goals, the countries and organizations involved, and the activities and progress thereof.

The expectations from the HGP, as held by the scientific, therapeutic and biotechnological communities, as well as society as a whole, are enormous, and so are the benefits (see Appendix 1).

At the same time, however, the knowledge derived from the HGP could, and most probably would, result in the establishment of a genetic identification for each person, harboring a great risk to the individual and to society if not used prudently and under the strictest of regulations. (See Appendix 2).

THE HUMAN GENOME PROJECT

EVOLVEMENT

1. Beginning of Human Genome Mapping: In the 1930s, following the discoveries that color blindness and hemophilia were linked to chromosome X.
2. Genetic mapping accelerated in pace in the late 1970s with the advent of RFLP (restriction fragment length polymorphism) mar kers. But these are relatively rare and unevenly dispersed in the Human Genome, difficult to analyze and not sufficiently informa tive.
3. Genetic mapping was boosted with the introduction of STRP (short tandem repeat polymorphism) markers in the late 1980s (first by Jim Weber, Marshfield Medical Research Foundation, CA).
4. With the aid of STRPs, the entire genetic map has been saturated by markers, so much so that the new maps incorporate over 3,600 STRPs, 400 genes, and 1,800 other markers (RFLPs and other DNA segments). These maps describe human genetic diversity at a mean resolution of 0.7 cM (centimorgan).
5. Genetic maps helped localize more than 40 genes, including cystic fibrosis, fragile X syndrome, myotic dystrophy, types of colon and breast cancer (BRCA1), ataxia telangiectasia, Alzheimer's disease, and others. Application of gene therapy has been in progress since 1990 (see Appendix 1).
6. The short-term goal of genetic mapping has been accomplished, but it leaves too many gaps and lacks anchor points of chromosomal telomeres and centromere. A satisfactory genetic map is believed to be achieved by increasing the marker density up to 1 marker per 100 bp.
7. Today (1st half of 1995), about only 2.5% of the Human Genome has been sequenced.

THE PROCESS OF DECIPHERING THE HUMAN GENOME

A. EXPERIMENTAL PROCEDURES:

1. **RFLPs** (introduced by Solomon and Bodmer in 1979, and Botstein and co-workers in 1980). Since DNA varies from one individual to another with roughly 1 nucleotide per 500, when DNA is cut with restriction enzymes a polymorphic pattern of fragments is produced in different individuals, which can be employed in genetic mapping by finding RFLPs with similar traits (markers).
2. **Pulsed-field gel electrophoresis (PFGE)** (Schwarts and Cantor, in 1984) enables separation of large DNA fragments up to 10 M bp.
3. Polymerase chain reaction (PCR) (Saiki and co-workers, in 1985, and Mullis and coworkers, in 1986) enables a manifold amplification of a DNA sequence, providing working means for analyzing minute amounts of DNA.
4. **Yeast artificial chromosome (YAC)** (Burke and coworkers, in 1987) enables cloning of large DNA segments up to 1 M bp.
5. **Sequence-tagged site (STS)** (Olson and coworkers, in 1989), the common mapping language, is a short, 100-1000 bp DNA segment, unique in the genome, and defined by a pair of PCR primers. "Genomatron" is an automated system that can screen hundreds of STSs in hours (developed by Eric Lander and co-workers, Whitehead Institute, 1994).
6. **Positional cloning:** "Positional candidate" strategy is predicted to become the major process for identifying disease genes. This approach is based on a 3-step process that saves time and effort: a) localizing a disease gene to a chromosomal subregion (using the traditional linkage analysis), b) searching databases for an attractive candidate gene within that subregion, c) testing the candidate gene for disease-causing mutations. It is believed that by now, the first quarter of 1995, it helped in identifying more than 50 disease genes.

B. THE EMBARKMENT OF THE HUMAN GENOME PROJECT:

1. The Human Genome Project (HGP) originated at the DOE Meeting in Alta, Utah in December 1984, where the possible use of DNA analysis in detecting mutations among atomic bomb survivors was contemplated.
2. This was then followed by the strive for sequencing of the entire Human Genome, advocated by several scientists, including Robert Sinsheimer (1985, then chancellor of the University of California, Santa Cruz), Charles Delisi, DOE, who described the framework (The Human Genome Project, Am Scientist 76, 488-493, 1988), and Renate Dulbecco (1986, president of the Salk Institute).
3. A National Research Council (NRC) Committee was asked, in September 1986, to determine whether the Human Genome Project (i.e., sequencing the Human Genome) should be advanced. In February 1988 the Committee recommended the implementation of the HGP, which will include in addition to Human Genome mapping also mapping of model organisms, at a budget of $200 m per year for 15 years, in which NIH will have a central role.
4. Another committee, appointed by the US Congress' Office of Technology Assessment (OTA), released a report in April 1988 supporting the recommendation of the NRC Committee.
5. In 1988 the Congress appropriated $17.3 m to NIH and $11.8 m to DOE for genome research.
6. In March 1988, James Wyngaarden, then NIH director, announced the creation of an NTH Office Center for Human Genome Research.
7. James Watson was appointed in October 1989 to head the office which became the National Center for Human Genome Research (NCHGR), serving as its director until April 1992; Michael Gottesman served as acting director until Francis Collins became the second and present director.
8. NIH and DOE, working as partners in managing HGP, presented to Congress a 5-year-term program in early 1990 with 8 major goals:
   a. Develop maps of human chromosomes.
   b. Improve technology for DNA sequencing.
   c. Mapping and sequencing DNA of selected model organisms

(mouse, *Caenorhabditis elegans, Drosophila melanogaster, Saccharomyces cervisiae, Escherichia coli).*

d. Bioinformatics (collect, manage and distribute data).
e. Study the legal, social and ethical issues, and develop policy options.
f. Develop and improve technology.
g. Facilitate the transfer of technology.

Work is coordinated by a joint NIH-DOE Subcommittee on Human Genome.

9. Bioinformatics: A number of databases has been created, such as the Genome Data Base (GDB) supported by NIH and DOE at the John Hopkins University Welch Medical Library.
10. ELSI, a program for the ethical, legal and social implications, has been launched.
11. Progress for the 1st 5-year period is right on schedule (see also Fig. 1), especially genetic mapping and sequencing of model organisms, while sequencing techniques are being improved progressively. The results of the linkage map are being published in the "Index Marker Catalog" of NCHGR, and complete mapping with 10-15 cM resolution was completed in 1993.

NATIONAL AND INTERNATIONAL ACTIVITIES IN THE HGP

The United States leads in terms of effort, cost support and results, followed by France, the UK, and to a lesser extent Japan. Specific programs in HGP have been introduced in other countries, including: Australia, Belgium, Canada, Denmark, Germany, Israel, Italy, The Netherlands, Russia and Sweden. In addition, there is a growing international collaboration of international organizations involved in the HGP, especially in dissemination of information and informatics and in mapping and sequencing model organisms (see below). Three major international organizations are involved in the HGP:

I. HUGO
II. EC
III. UNESCO

I. HUGO is involved primarily in dissemination of knowledge in various forms {informatics (data banks), information, meetings and workshops}.
II. EC is engaged in processing various functions via its framework programs of science and biotechnology.
III. UNESCO, via its International Bioethics Committee, examines the legal, social and ethical aspects of HGP, at the international level.

The support for the activities in HGP comes primarily from the governments, and in some, most notably the UK and France, from charities (see below).

1. THE UNITED STATES

In the United States, the Human Genome Project (HGP) was initiated at DOE in 1986, and at the NTH a year later, in 1987. The HGP is today coordinated by NIH and DOE at a budget of $200 m per year, to continue until 2005, at an overall expenditure of $3 b ($3 \times 10^9$).

In the U.S. there are 21 centers engaged in this project (see Bibliography No. 1) that are working on:

a. Physical mapping of human chromosomes 3,4,5, 6,11,12,13,15, 16, 17, 19, 21, X.
b. Physical mapping of mouse chromosomes 11 and X.
c. Complete mapping of the *Drosophila melanogaster* genome.
d. Mapping of *Saccharomyces cervisiae* chromosomes IV and V as part of an international effort to sequence the entire yeast genome by 1996.
e. The entire genome of *E. coli* K12 strain MG 1655 (4.7 kb).
f. The entire genome of *E. coli.*
g. The complete genomic sequence of the nematode *Caenorhabditis elegans,* in collaboration with the Sanger Center, UK.
h. The complete sequence of 2 bacteria: *Haemophilus influenzae* and *Mycoplasma genitalium* (published in May 1995).

The NIH National Center for Human Genome Research (Francis Collins, director) and DOE are examining the means by which the beneficial therapeutic aspects of the Human Genome Project, along with the legal, social and ethics problems that it embodies, can be consummated.

2. FRANCE

a. **Combined effort of two centers.** Genethon - £6.5 m, an automated laboratory facility supported by the French Muscular Dystrophy Association, and Centre d'Etude Polymorphisme Humain, have placed France second to the United States in the effort in Human Genome Research. In 1992, Genethon completed mapping 28% of the human chromosome 21 and provided about 2,000 genetic markers. In 1993 it published a first-generation map (low resolution) of 90% of the Human Genome.
b. **Human genetic maps** (CEPH: Centre d'Etude du Polymor phisme, in collaboration with the United States). The program provides collaborative investigators with DNA from cultured lymphoblast cell lines derived from a reference panel of more than 60 nuclear families from France, Venezuela, Pennsylvania, and Utah.
c. **Financial support:**

I. Government (CEPH)
II. "Genethon": Project funded by the Muscular Dystrophy Association
IH. Television sponsorship appeal (FF300 m was raised in one night).
IV. Development of the largest center in the world for gene analysis, automated processing with additional automated machines from the EUREKA Labimap 2001 Project.
The aims of the center are to:
I. Collect blood samples from 25,000 disease families.
II. Map the disease genes.

d. Current (1990) overall support: £15.6 m
e. In 1988, the Ministry of Research and Technology funded a Genome Concerted Actions Programme, which will also support the EUREKA Labimap 2001 Project.
f. In the autumn of 1990, the Minister of Research and Technology launched the French Human Genome Research Programme, whose main task is to coordinate the efforts of the different bodies, including the automation of molecular biological methods (e.g., through the EUREKA Labimap 2001 Programme).
g. **Projects:** Human Chromosome 20 (genetic mapping), in progress at the Institut Pasteur, will be transferred to Genethon Centre; other human chromosomes will possibly be included later.
h. **The EUREKA Labimap 2001 Genome Project,** initiated in 1988 for 4 years, is run jointly by France and the UK, involving CEPH, UK Imperial Cancer Research Fund, and Amersham Int., coordinated by Bertin et Cie. Its aim is the development and commercialization of automated sequencing machines.
i. **Budget** (1990,1991) of the Ministry of Research and Technology.
1990, FF7 m - Genome Concerted Actions Programme
1990, FF30 m - EUREKA Labimap 2001 Project
1991, FF50 m - new Human Genome Programme
1992 onwards, FF100 m - new Human Genome Programme, j.
Model organisms included Mouse genome; *E. coli* and *Bacillus subtilis, C. elegans,* and the plant *Arabidopsis thaliana.* k.

International collaboration: At the Cold Spring Harbor meeting

in May 1995, Jean Morisette of Genethon presented an almost complete human genetic linkage map, covering the entire human genome with 5,300 markers (2.5 times the number available 18 months ago), and Isabelle le Gall of the Centre d'Etude du Polymorphisme Humain (CEPH) in Paris presented an improved physical map, comprised of overlapping sequences of mega-YACs, covering up to 75% of the human genome with 33,000 YACs. For detailed information on activities in France see Bibliography No. 2.

3. UNITED KINGDOM

a. Most of the work in the Human Genome Mapping Project (HGMP) is conducted in academic institutions, in specific genome centers, as well as in university departments. Some commercial activity is beginning to surface.
b. Support:

I. Government
II. Wellcome Trust
III. Imperial Cancer Research Fund

I. **Government:** The support for HGMP, via the Medical Research Council (MRC) and HGMP, is reviewed by an internal committee. In 1989, £11 m was allocated for a 3-year-period, and in 1992, £4.5 m per annum.

An HGMP Coordinating Committee orchestrates the activity within the MRC. This committee includes also representatives of interested charities and other research councils to expedite process.

An Industrial Advisory Group was established to interact with HGMP activities and committees. In 1993, an additional £1.5 m for one year was allocated to promote work on comparative mapping.

The MRC HGMP Resource Centre (founded at the Clinical Research Centre at Northwick Park Hospital), supported by the MRC, serves as a bioinformatics centre for the human and

mouse genome. Services offered are free for scientists in return for data deposition, while industry pays a small fee to cover costs. In 1994, the Centre moved to Hinxton Hall adjacent to the MRC/Wellcome funded Sanger Centre.

Agriculture and Food Research Council: The Biotechnology and Biological Sciences Research Council is involved in identifying genes of commercial value (such as growth, disease resistance, etc.), and mapping of several species, including: pig (as part of the European PigMap effort), limited effort on chicken and cattle, and plants *(Arabidopsis thaliana* - model, and also pea, wheat, barley, grasses and oilseed rape).

II. **The Wellcome Trust:** Funded the establishment of two key centers:
   i) The Sanger Centre in Cambridge (physical mapping and sequencing of Human Genome, the worm Caenorhabditis elegans, and yeasts).
   ii) In Oxford, a £13 m, 5-year study on genetic diseases. To oversee these activities, the Wellcome Trust established a Genetic Interest Group to coordinate activities and funding.

III. **The Imperial Cancer Research Fund:** Made a major contribu tion in establishing clones (YAC libraries as well as cosmid, cDNA and P1 libraries) for human, mouse and yeast genomes, and supports the development of sequencing technology.

c. **Species evaluated:** Primarily, mouse and C. *elegans,* also human, pig, chicken, cattle, yeast, plants (Arabidopsis thaliana).
d. **The UK's strength** lies in the robotics supported cloned DNAs immobilized on filters. The UK is the leader in throughput and the density of clones on the filters.
e. **International activities:** The Sanger Centre serves as an Interna tional EST (expressed sequence tagged) and includes consortium, Stanford University Genome Center, Wellcome Trust Centre for Human Genetics, Genethon, Washington University, University of Cambridge, and Whitehead Institute at MIT. The aim is to map 70,000 ESTs at intervals of 0.5 Mb or less. To facilitate this

process, the Institute of Genomic Research donated primers for 15,000 ESTs.

4. JAPAN

Less awareness of and exposure to genetic diseases.

a. **Informatics:** Science and Technology Agency (STA):
   I. Gene Data Base: Operating under STA, at the Japan Information Center of Science and Technology (1993).
   II. **DNA Data Bank of Japan (DDBY):** established at the National Institute of Genetics in 1986. Connected to GenBank (United States) and EMBL (Europe).
   III. **Genome Net:** Inter-university network within Japan.

b. **The Human Genome Center:** Established at the University of Tokyo Medical Research Institute, it is dedicated to promoting research.

c. **Support:** Several Government ministries including: Ministry of Science and Technology Agency (STA) and Ministry of Educa tion, Ministry of Science and Culture (Monbusho), and to a lesser extent, the Ministry of Health and Welfare and Ministry of Agriculture, Forestry and Fisheries (the latter because of projection of HGP to other organisms, including rice).
   I. Started in 1981, the development of automated sequencing machines
   II. Since 1988, £7 m, half of which has been used for automated methods.

   Overall support (1991):

   £1.2 m - The Ministry of Education Science and Culture MESC program

   £3.7 m - The Science and Technology (STA) WADA (automated sequencing machine) project

   £0.3 m - Silver Science Program Yoken Project

   £2.4 m - The Ministry of Agriculture, Forestry and Fisheries (MAFF) Rice Genome Project

   Total: £7.6 m

d. **Model organisms:**
E. coli (started 1987, 3 year program, £1.4 m). Also, on a smaller scale: *B. subtilis, S. cerevisiae, Saccharomyces pombe, C. elegans, Arabidopsis thaliana.* Plant chloroplasts (1984-1988, £850 k; sequenced 3 chloroplasts of 420 k bp).

5. **ISRAEL**

The major body directing the scientific activities pertinent to the Human Genome Project (HGP) in Israel is the Israel Academy of Sciences and Humanities (Academy), assisted by its Advisory Committee on the Human Genome which was nominated in March 1991.

Within its capacity of promoting the advancement of basic research in Israel, the Academy has been engaged in invigorating activities in the HGP along four different lines:

i) the establishment of national centers;
ii) the support and stimulation of research by earmarked grants;
iii) the attraction of young scientists to the field through post-doctoral grants; and
iv) the dissemination of knowledge through meetings and workshops.

a. **HGP Centers**
In 1993 two national centers, whose major task is to assist Israeli researchers in the Human Genome field, was launched: the Bioinformatics Genome Resource Center at the Weizmann Institute of Science in Rehovot, and the National Laboratory for the Genetics of Israel Populations at Tel Aviv University.

I. The Bioinformatics Center of the Weizmann Institute of Science serves as the Israel National Node within the EMBL's International Nodes network, and is shortly due to connect to the European Bioinformatics Institute in Cambridge, England.

In this capacity, the Bioinformatics Center provides scientists in Israel with access to the vast and rapidly growing genetic information (DNA maps and sequencing, and protein

structure) produced in the world, primarily that derived from the HGP.

II. **The National Laboratory for the Genetics of Israel Populations** at Tel Aviv University will also serve as one of the regional centers, the Israel International Center for Human Genetic Diversity, within the International Project of "Human Gene Diversity". Israel harbors a unique advantage owing to the vast immigration from many countries of large genetic diversity.
The Center is engaged in blood sampling and preserving (by immortalization) of cell lines of various Israeli populations, and hopefully other neighboring population groups in the future. The material stored in the Center will be available for qualified scientists, national and international, to study the diverse genetic traits (normal and pathologic) for the advancement of knowledge in various fields, including anthropology, biology and medicine. Being aware of the potential harmful risks involved, as a result of misuse of the vast information accumulated there, the Center will exercise appropriate precautions and protective measures to assure the protection and well-being of the donors.

b. **Major activities**

I. Chromosome 17. The main contribution and activity of Israel within the HGP is at the human chromosome 17. Prof. Doron Lancet of the Weizmann Institute of Science was recently nominated by HUGO to serve as 'editor' of the Human Chromosome 17.

II. Study of genetic traits (normal and pathologic) of Jewish populations, begun about 30 years ago, has been boosted by the establishment of the National Laboratory for the Genetics of Israel Populations at Tel Aviv University.

III. Cloning, mapping, sequencing and evaluating several hu man genes, primarily of pathological implications and with Jewish orientation, such as ataxia telangiectasia, cystic fibrosis, Gaucher's disease, Tay-Sachs disease, to name a few.

c. **Expenditures** for the various activities in the HGP for the 1995

financial year exceeds $600 k. Supported by the government, through the Planning and Budgeting Committee of the Council for Higher Education, and by several charities.

6. INTERNATIONAL ORGANISATIONS

a. HUGO (The Human Genome Organization)

I. **Inception:** Established in April 1988, by 42 scientists from 17 countries at the Cold Spring Harbor Meeting.

- September 1988, the Founding Council convened, headed by a president and 3 vice-presidents and a director.
- July 1989, institution of the HUGO Headquarters in Geneva, and 3 permanent regional offices: America Office in Bethesda, Europe Office in London, and the Pacific Office in Osaka.

II. **Funding:** Initially by several charity organizations, including Howard Hughes Medical Institute (HHMI), Wesley Foundation, Lucille P. Markey Charitable Trust, and the Imperial Cancer Research Fund.
For example, in 1990: HHMI - $1 m, Wellcome Trust - £175 k.

III. **Administration activities:** including that of the president and the director, are covered by the governments (primarily the UK).

IV. Membership: "Membership shall be open to all persons concerned with the human genome or other scientific subjects related to it", by annual election. No membership fee.

V. **Functions:** Coordinates exchange of knowledge and data, including seminars, workshops and training, and encourages public awareness of the scientific, medical, technological, legal, social and ethical implications of the HGP. These functions are promoted via committees:
- Human Genome Mapping Workshops Committee
- Physical Mapping Committee
- Informatics Committee
- Mouse Mapping Committee
- Committee on Ethical, Social and Legal Issues
- Committee on Intellectual Property and Ownership

b. EMBL (the European Molecular Biology Laboratory)

Located in Hamburg and funded by some, but not all member countries. Focuses on two areas:

- Database coordination (informatics)
- Development of instrumentations and techniques

DATABASE ACTIVITIES

I. **The Nucleotide Sequence Data Library** (United States GenBank is about 50% larger).

n. **EMBNET:** Established in 1989, it is comprised of nodes (in 1991 there were 15 nodes), one of which is the Israeli node at The Weizmann Institute. It is connected to GenBank and the Data Bank of Japan.

HI. **Swissport:** A collection of amino acid sequences along with translations of coding sequences in the EMBL Nucleotide Sequence Database, developed at the University of Geneva and maintained by EMBL. Also connected to the protein sequence database in the United States.

c. **UNESCO (United Nation Educational, Scientific and Cultural Organization):** Program for 1990/91, with a budget of $350 k.

Main functions:

I. Dissemination of information.

II. Ethical and social issues (International Bioethics Committee, since 1993).

d. **Informatics:**

Three main centers:

I. EMBL in Heidelberg

II. NCBI (National Center for Biotechnology Information) in Washington DC, for sequences.

III. GDB (Genome Database) in Baltimore, human genetic maps.

INFORMATIVE DATA, TERMINOLOGY AND METHODS

1. The Human Genome is comprised of 23 pairs of chromosomes - 22 autosomes and 1 sex chromosome (X,Y) having $3x10^9$ bp of DNA, about 100,000 genes, of which about 5,000 are disease genes.
2. The smallest human chromosome: Y, 50 M bp. The largest human chromosome: 1, 250 M bp. Average-sized gene: 30 K bp, encoding a 1,000 amino acid protein.
3. Karyotype: analysis of chromosomes via microscope based on shape (size and banding pattern).
4. **Mapping and sequencing the Human Genome**
   Mapping: dividing each chromosome into small segments, characterizing them, and arranging them sequentially (mapping) on the chromosome. A genome map describes the order of genes, other known DNA segments of no known functional protein, and the spacing between them on each chromosome.
   a) **Genetic Map** depicts the order by which genes are arranged along a chromosome. The determination of such a sequence (Genetic Map) is facilitated by known markers: genes or other DNA stretches. Distances between markers are measured in centimorgans (cM). 1 genetic cM is about 1 M bp (on physical distance).
   b) **Genetic Linkage Map:** Shows the relative location of a specific DNA marker along the chromosome.

   Markers must be polymorphic (variations in DNA sequence occurring once every 300-500 bp) to be useful in mapping. Most variations occur in introns, whereas in exons this could result in observable changes such as eye color, etc.
   The human genetic linkage map is constructed by observing how frequently two markers are 'interherited' together. The closer the markers are to one another on the same chromosome, the more tightly linked they are, the more likely they will be passed together to the next generation, i.e., they will not be separated by recombination events. Hence, the distance between two markers

can be determined. This can also assist in locating a gene, especially that of genetic disease on a chromosome. Genetic maps assisted in chromosomal location of several inherited diseases, such as sickle cell disease, cystic fibrosis, Tay-Sachs disease, fragile X syndrome, myotonic dystrophy, ataxia telangiectasia.

Goals: Complete detailed genetic map of 1 cM resolution.

c) **Physical Map** shows the actual sites of genes on the genome. There are several physical maps with different degrees of resolution. The physical map of DNA, like a topographic map, is comprised of mapped landmarks, such as restriction enzymes and STS (see below), providing reference points relative to which functional DNA sequences such as genes can be localized.

i. The lowest resolution physical map is the Cytogenetic Map where chromosomal band patterns are viewed with light microscope of stained chromosomes.
ii. cDNA Map shows the location of genes.
iii. iii. The highest resolution map shows a complete DNA bp sequence.

Mapping Methods

a) Terminology

I. **RFLP** (restriction fragment length polymorphism): sequence variations in DNA sites that can be cleaved by restriction enzymes.
II. **STRP** (short tandem repeat polymorphism): variable number of tandem repeat sequences, most commonly of 2 bp.

The advantages of STRPs are:
- Repeated up to thousands of times throughout the Genome.
- Even distribution throughout the Genome.
- Amplifiable by PCR.
- Number of repeats vary among individuals.

In some cases, an excess of repeats of trinucleotides could lead to

inherited diseases such as fragile X syndrome, Huntington's disease, myotonic dystrophy.

III. **STS** (sequence-tagged sites): A short, unique sequence of DNA that can be amplified by PCR. Ideal landmarks during map construction (easily detectable by PCR). The most critical aspect of an STS description is the DNA sequence of the 2 primers.

IV. **Contig** (Contiguous = sharing edge or touching with over-lapping regions of a genome): An organized set of DNA clones that collectively provide redundant cloned coverage of a region too long to clone in one piece.

V. **YAC** (yeast artificial chromosome): A cloning system, up to 1 Mb DNA segment.

VI. **Cosmid:** 20-40 kb DNA fragment.
Both YAC and Cosmid are used for cloning, enabling the use of these clones for sequencing (fragmenting) by several groups.

VII. **FISH** (fluorescence in situ hybridization): A physical mapping technique employing fluorescin-labeled DNA probes that can detect segments of the human genome by DNA-DNA hybridization on samples of condensed chromo somes of lysed metaphase cells.

VIII. **Centimorgan** (cM): A unit of measure of recombination frequency. 1 cM represents a 1% chance that a marker at one genetic locus will be separated from a marker at a second locus as a consequence of crossing over in a single generation. In humans, 1 cM is about 1M bp.

b) **Methods**

I. **Macrorestriction Maps:** Top-down mapping: Fragmenting chromosomes with a rare restriction enzyme into large pieces, which are then ordered and subdivided and then mapped. This result in more continuity and less gap than the contig method, but it has a lower map resolution.

II. **Contig map** (Bottom-up mapping): Cutting a chromosome

into small pieces, each cloned and ordered, forming contiguous DNA blocks.

III. **Positional Cloning:** The markers are used for gene hunt. Once the gene is located, physical maps are used to obtain flanking DNA segments for further detailed study (mostly pertains to regulation of gene function).

IV. **STS-Content Mapping:** Provides the means to establish these overlaps between each clone and its nearest neighbors. If 2 clones share even a single STS, they can reliably be assumed to overlap.
Using the YAC system and STS-content mapping, physical maps of human chromosomes 21 & Y, and a large part of X have been published. {YAC gives a much larger segment (contigs) than those observed with cosmid clones.} The problem with YAC (and more so with cosmid) is the orientation of the contig obtained. This is overcome by the FISH method, where a fluorescence-labeled probe that binds to a chromosome is visualized through light microscope.

V. **Radiation Hybrid Mapping:** Involves fragmentation of chromosomes in cultured cells with high doses of X-rays followed by incorporation of the fragments into stable cell lines.

VI. **PCR** Markers: Based on short, repetitive DNA sequences widely distributed in the human genome, such as: (CA)n (n, number of repetitions of the dinucleotide CA). n is highly variable in the different zones of the Human Genome. The difference in n results in different copy length, detectable by electrophoresis.

VII. **Positional Cloning:** Identifying a gene inflicting a disease (inheritable disease).

The major technical hindrance in sequencing is that no sequencing machine(s) exists to provide a contiguous, finished DNA sequence on a large scale (> 1 Mb). This is a priority goal in the Human Genome Project.

GLOSSARY

Autoradiography: A technique that uses X-ray film to visualize radioactively labeled molecules or fragment of molecules. It is used to analyze the length and number of DNA molecules after they have been separated by gel electrophoresis.
Centimorgan: A unit of measure of recombination frequency. One centimorgan is equal to a 1% chance that a genetic locus will be separated from a marker due to recombination in a single generation. In humans, a centimorgan equals, on average, one million base pairs.
CEPM: Centre d'Etude du Polymorphisme.
cDNA: DNA that is synthesized from a messenger RNA template; the single strand form is used as a probe in physical mapping.
Chromosome: A structure of genetic information composed of protein and DNA. Each human cell has 46 chromosomes in pairs. One chromosome of each pair is inherited from each parent.
Contigs: Groups of clones representing overlapping or contiguous regions in a genome.
Cosmid vectors: Plasmids that also contain specific sequences from the bacterial phage lambda. Cosmids are designed for cloning large fragments (typically 40 kbp) of eukaryotic DNA.
Database: Repository of information.
DNA: DNA is normally a double stranded molecule that encodes genetic information, the two strands being held together by chemical bonds between base pairs of nucleotides. There are four such nucleotides, adenosine, guanosine, cytidine, and thymidine. Base pairs are formed only between adenosine and thymidine and between guanosine and cytidine. It is possible to determine the sequence of either strand from that of its partner.
DNA cloning: A means of isolating individual fragments from a mixture and multiplying each to produce sufficient material for further analysis.
DNA sequence: The relative order of base pairs in a stretch of DNA; of a gene, a chromosome or an entire genome.
EC: Abbreviation for the Commission of the European Communities.
Electrophoresis: A method used to separate and characterize large DNA

and protein fragments from a mixture of such molecules.
ELSI: Ethical, legal and social implications. A central issue within the HGP, and conducted by DOE.
EMBL: European Molecular Biology Laboratory, engaged in promoting HGP-related informatics, methods and techniques. EST (expressed sequence tagged): Short DNA sequences derived from active genes.
Eukaryote: All organisms except viruses, bacteria and blue-green algae.
Exons: The protein-coding DNA sequences of a gene. **GDB:** Genome Data Base.
Gene: The fundamental unit of inheritance; an ordered sequence of nucleotides that specify the manufacture of a single type of protein (or for some genes, certain RNAs).
Gene therapy: Insertion of normal DNA into cells to correct a given genetic defect.
Genetic linkage map: A map showing the relative positions of gene loci on a chromosome. The distance is measured in centimorgans.
Genetics: The study of patterns of inheritance of specific traits.
Genome: All the genetic material in all the chromosomes of a particular organism. Size is usually denoted in base pairs.
Genomic library: A collection of clones made up of a set of overlapping DNA fragments representing the entire genome. **Germinal cell:** A cell destined to become a sex cell. **HGMP:** Human Genome Mapping Project.
HUGO (Human Genome Organization): An international organization active in promoting HGP-related international cooperation and dissemination of information.
Hybridization: The joining of two complementary strands of DNA, or of DNA and RNA, to form a double stranded molecule.
Informatics: The application of computer and statistical techniques to the management of information.
Introns: The DNA sequences separating the protein coding sequences of the genes. Introns are transcribed into mRNA but are eliminated from the message before it is translated into protein.

Library: A collection of unordered clones whose relationship may be established by physical mapping.
Linkage: The proximity of two or more markers on a chromosome. The closer together the markers are the lower the probability that they will be separated during meiosis. This gives an idea of the likelihood that they will be inherited together.
Locus: The location of a gene on a chromosome.
Marker: An identifiable physical location on a chromosome whose inheritance can be monitored.
Nucleotide: The building block of DNA or RNA; thousands of nucleotides are linked to form a DNA or RNA molecule.
Oligonucleotide probe: A short DNA probe whose hybridization is sensitive to a single base mismatch.
Physical map: A map showing the identifiable landmarks (genes, RFLP markers, etc.) on DNA, regardless of inheritance. Distance is measured in base pairs. Chromosome banding patterns constitute the lowest resolution physical map; complete nucleotide sequences represent the highest resolution map.
Plasmid vectors: Circular, double stranded DNA of 1 k to 10 kbp that can carry additional DNA sequences in fragment inserts up to 12 k base pairs; vastly employed in genetic engineering.
Polygenic disorders: Genetic disorders resulting from the combined action of alleles of more than one gene; e.g., heart disease or diabetes. The hereditary patterns of these disorders are more complex than single gene disorders.
Polymerase chain reaction (PCR): A technique that allows a sequence of interest to be amplified selectively against a background of an excess of irrelevant DNA, up to 10^6 to 10^9 fold.
Polymorphism: Differences in DNA sequence among individuals. Genetic variations occur in more than one percent of a given population, and could be used for linkage analysis.
Pulsed-field gel electrophoresis (PFGE): A method used to separate large DNA fragments (20 kb to 10 Mb) by applying pulses of current to the sample at various angles.
Replication: The synthesis of new DNA strands from existing DNA.
Restriction enzyme: A protein that recognizes specific short nucleotide

sequences and cuts DNA at such sites.

Restriction fragment length polymorphism (RFLP): Polymorphic variants in the size of the fragments produced by digesting DNA with a restriction enzyme.

Ribonucleic acid (RNA): Similar in structure to DNA, RNA is comprised of uracil rather than thymidine nucleotide. Usually a linear single stranded polymer, encoding protein synthesis.

Sequence-tagged site (STS): A short DNA sequence, readily located and amplified by PCR techniques, that uniquely identifies a physical genomic location.

Sequencing: Ordering of nucleotides.

Single gene disorders: Hereditary disorders caused by a single gene; e.g., Duchenne muscular dystrophy.

Somatic cell: Any cell of the eukaryotic body other than those destined to become sex cells.

Telomere: DNA that forms the ends of chromosomes and which is needed for successful replication.

Trait Any genetic characteristic.

Vector: A DNA molecule capable of autonomous replication in a cell and which contains restriction enzyme cleavage sites for the insertion of foreign DNA.

Whole-genome radiation hybrid mapping (David Cox, Stanford Univ; G. Gyapay, Genethon): A complementary technology for constructing highly integrated maps of human chromosomes.

Yeast artificial chromosome (YAC): Plasmids containing portions of yeast chromosomal DNA that function in replication. Used to clone foreign DNA fragment inserts up to one million base pairs in length.

BIBLIOGRAPHY

1. Human Genome Research: A Review of European and International Contributions. Medical Research Council, Diane J. McLaren, ed., January 1991.

2. Report on Genome Research 1991. The European Science Foundation.

3. DOE Human Genome Program. Primer on Molecular Genetics. June 1992.

4. Maynard V. Olson. The Human Genome Project. PNAS, 90, 4338-44, 1993.

5. F. Collins and D. Galas. A New Five-Year Plan for the U.S. Human Genome Project. Science, 262, 43-46, 1993.

6. L.W. Engel. The Human Genome Project. Arch. Pathol. Lab. Med., 117, 459-465, 1993.

7. Human Genome News. NTH and DOE. Vol 6, No. 4, 1994.

8. The Human Genome Mapping Project in the UK. Priorities and Opportunities in Genome Research. Commissioned by the Office of Science and Technology, April 1994.

9. Human Genome 1993, Program Report. U.S. Department of Energy, March 1994.

10. R. Nowak. Cold Spring Harbor Meeting Report. Science, 268,1134-5, 1995.

11. The Genome Directory. Nature, Vol. 377, Issue No. 6547S, 1995.

APPENDIX 1

Gene Therapy

The primer benefit to be derived from the Human Genome Project is gene therapy, in which defective genes inflicting congenital diseases are replaced by functional genes.

Gene therapy was launched successfully in 1990 when the first approved human gene therapy was performed. Its recipient was Ashanthi De Silva, a girl aged 3, who suffered from congenital adenosine deaminase deficiency. If untreated, this deficiency could lead to a fatal malfunction of the immune system. The functional adenosine deaminase gene was successfully introduced.

Since then more than 100 clinical trials have been introduced, most with less success.

The last 5 years have witnessed an intensive effort, both in the public and private sectors, involving many hundreds of millions of dollars in the United States alone (about $200 M by NIH, surpassed by private companies), towards the development of tests and techniques for gene therapy. This 5-year experimentation period has helped to air the safety-related doubts with regard to possible hazards that engineering DNA might impose, such as the promotion of new infectious diseases and cancers.

However, although still considered a most rewarding prospect, gene therapy at present remains unclear, posing more questions than answers, such as the right vector to be used, side effects (noted in several cases of adenovirus use in cystic fibrosis trials), and, above all, its efficacy.

APPENDIX 2

THE DECLARATION OF INUYAMA

Human Genome Mapping, Genetic and Screening Therapy

The Council for International Organizations of Medical Sciences held its XXIVth, Round Table Conference, on the subject of *Genetics, Ethics and Human Values: Human Genome Mapping, Genetic Screening and Therapy,* in Tokyo and in Inuyama City, Japan, from 22-27 July 1990. The Conference was held under the auspices of the Science Council of Japan and cosponsored by the World Health Organization and the United Nations Educational, Scientific and Cultural Organization. It was the fifth in a series entitled *Health Policy, Ethics and Human Values: An International Dialogue,* begun in Athens in 1984. the participants, numbering 102, came from 24 countries, representing all continents.

In addition to biomedical scientists and physiciains, the participants represented a wide range of disciplines including sociology, psychology, epidemiology, law, social policy, philosophy and theology, and brought with them experience in hospital and public health medicine, universities and private industry, and the executive and legislative branches of government. Through presentations and discussions in plenary sessions and working groups, they reached broad agreement on a number of central issues. At its final session the Conference agreed on the following Declaration.

I. Discussion of human genetics is dominated today by the efforts now under way on an international basis to map and sequence the human genome. Such attention is warranted by the scale of the undertaking and its expected contribution to knowledge about human biology and disease. At the same time, the nature of the undertaking, concerned as it is with the basic elements of life, and the potential abuse of the new knowledge which the project will generate are giving rise to anxiety. The Conference agrees that efforts to map the human genome present no inherent ethical problems but are eminently worthwhile, especially as the knowledge revealed will be universally applicable to benefit human

health. In terms of ethics and human values, what must be assured are that the manner in which gene mapping efforts are implemented adheres to ethical standards of research and that the knowledge gained will be used appropriately, particularly in genetic screening and gene therapy.

II. Public concern about the growth of genetic knowledge stems in part from the misconception that while the knowledge reveals an essential aspect of humanness it also diminishes human beings by reducing them to mere base pairs of deoxyribonucleic acid (DNA). This misconception can be corrected by education of the public and open discussion, which should reassure the public that plans for the medical use of genetic findings and techniques will be made openly and responsibly.

III. Some types of genetic testing or treatment not yet in prospect could raise novel issues - for example, whether limits should be placed on DNA alterations in human germ cells, because such changes woulf affect future generations, whose consent cannot be obtained and whose best interests would be difficult to calculate. The Conference concludes, however, that for the most part present genetic research and services do not raise unique or even novel issues, although their connection to private matters such as reproduction and personal health and life prospects, and the rapidity of advances in genetic knowledge and technology, accentuate the need for ethical sensitivity in policy-making.

IV. It is primarily in regard to genetic testing that the human genome project gives rise to concern about ethics and human values. The identification, cloning and sequencing of new genes without first needing to know their protein products greatly expand the possible scope for screening and diagnostic tests. The central objective of genetic screening and diagnosis should always be to safeguard the welfare of the person tested: test results must always be protected against unconsented disclosure, confidentiality must be ensured at all costs, and adequate counselling must be provided. Physicians and others who counsel should endeavour to ensure that all those concerned understand the difference between being the carrier of a defective gene and having the corresponding genetic disease. In autosomal recessive conditions, the health of carriers, (heterozygotes) is usually not affected by their having a single copy of the disease gene; in dominant disorders, what is of

concern is the manifestation of the disease, not the mere presence of the defective gene, especially when years may elapse between the results of a genetic test and the manifestation of the disease.

V. The genome project will produce knowledge of relevance to human gene therapy, which will very soon be clinically applicable to a few rare but very burdensome recessive disorders. Alterations in somatic cells, which will affect only the DNA of the treated individual, should be evaluated like other innovative therapies. Particular attention by independent ethical review committees is necessary, especially when gene therapy involves children, as it will for many of the disorders in question. Interventions should be limited to conditions that cause significant disability and not employed merely to enhance or suppress cosmetic, behavioral or cognitive characteristics unrelated to any recognized human disease.

VI. The modification of human germ cells for therapeutic or preventive purposes would be technically much more difficult than that of somatic cells and is not at present in prospect. Such therapy might, however, be the only means of treating certain conditions, so continued discussion of both its technical and its ethical aspects is essential. Before germ-line therapy is undertaken, its safety must be very well established, for changes in germ cells would affect the descendants of patients.

VII. Genetic researchers and therapists have a strong responsibility to ensure that the techniques they develop are used ethically. By insisting on truly voluntary programmes designed to benefit directly those involved, they can ensure that no precedents are set for eugenic programmes or other misuse of the techniques by the State or by private parties. One means of ensuring the setting and observance of ethical standards is continuous multidisciplinary and transcultural dialogue.

VIII. The needs of developing countries should receive special attention, to ensure that they obtain their due share of the benefits that ensue from the human genome project. In particular, methods and techniques of testing and therapy that are affordable and easily accessible to the populations of such countries should be developed and disseminated whenever possible.

APPENDIX 3

GENETIC ENGINEERING

Genetic engineering will serve as the major therapeutic means in correcting gene mutation of genetic diseases.

Of the many techniques available several seem to be promising:
Transfection: Introduction of new genes.
Homologous recombination: Curing certain mutations in situ.
Injection of new genes into the nuclei of single-cell embryos.
Retroviral vectors: Introducing useful gene sequences into defective cells.

There are four potential areas for the application of genetic engineering designed to insert a gene into a human:

Somatic cell therapy: This would result in correcting a genetic defect in the body cells of a patient.

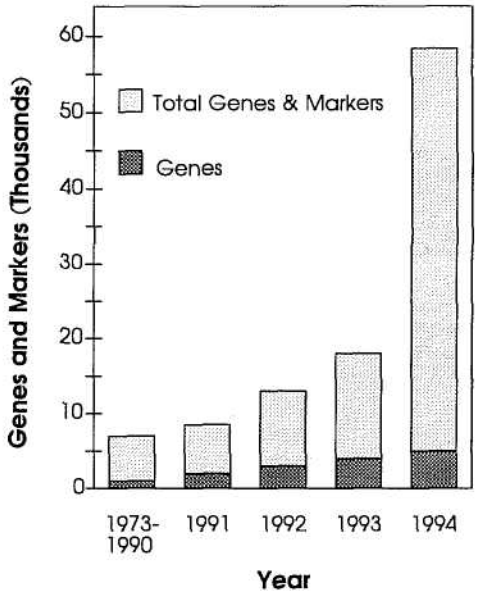
Germ-line gene therapy: This would require the insertion of the gene into the reproductive cells of the patient so that the disorder would also be corrected in future generations.

Enhancement genetic engineering: This would involve the insertion of a gene to enhance a known characteristic of a person, such as placing an additional growth hormone gene into an otherwise normal child.

Eugenic genetic engineering: This would represent the attempt to alter or improve complex human traits which are coded by a large number of genes, involving, for example, personality, intelligence, and formation of body organs.

- Somatic cell gene therapy is regarded technically as the simplest and ethically the least controversial form of gene therapy.
- Germ-line gene therapy will require major advances in present knowledge, and also raises ethical issues that clearly need to be debated further.

- Enhancement genetic engineering presents a host of difficult ethical concerns. Unless this type of therapy can be clearly justified on the grounds of preventive medicine, prevailing opinion suggests that enhancement engineering should not be performed.
- Eugenic genetic engineering is still prohibited and will probably remain so for the foreseeable future despite all the attention it continuously receives in the public and political arena.



The figure depicts the rapid growth of mapping data from 1973, when 64 genes were reported, through 1994 (as of 10/28/94), when over 40,000 genes and markers were submitted to the Genome Data Base (GDB), Nearly 60,000 genes and markers have been collected in all. Types of markers include D-segments - arbitrary DNA segments of unknown function - such as RFLPs and STRPs. Before 1991, all mapping data were gathered at Human Gene Mapping Workshops; data from 1991-94 were submitted to GDB. (Graph data provided by GDB.)

Figure 1. from the Human Genome News, sponsored by the National Institutes of Health and the U.S. Department of Energy, Volume 6, Number 4, p. 14, November 1994. With permission from the Human Genome Management Information System, Oak Ridge National Laboratory, 1060 Commerce Park, MS 6480, Oak Ridge, TN 37830.