Iatriki 1999, 76(3):256–263

Progress in the human genome project

O.A. Giotakos
Institute of Psychiatry, Section of Genetics, London, England

ABSTRACT *During the last years, a worldwide research effort has the goal of analyzing the structure of human DNA and determining the location of human genes. Called the human genome project, the effort got under way in 1990; it is a 15-year effort to identify all 80,000 genes in the human genome, and to determine the sequence of the 3 billion chemical bases that make up the human DNA. Mapping involves dividing the chromosomes into smaller fragments and placing them to their respective locations on the chromosomes. Many genetic and physical maps have been constructed; the most complete map was published in the summer of 1997 and featured about 8,000 landmarks. Geneticists have already charted the approximate position of over 2,300 genes, and roughly 2.5% of the human genome has been sequenced so far. The ultimate goal of the genome research is, to identify all the genes in the DNA sequence and to develop tools for using this information in the study of human biology and medicine. It is estimated the completion of the human genome sequence will take place by the end of 2003.*

Key words Human genome project, genetic map, physical map, sequencing.

Ιατρική 1999, 76(3):256–263

Η πορεία της έρευνας για το ανθρώπινο γονιδίωμα

Ο.Α. Γιωτάκος
Ινστιτούτο Ψυχιατρικής, Τομέας Γενετικής, Λονδίνο, Αγγλία

ΠΕΡΙΛΗΨΗ *Τα τελευταία χρόνια γίνεται μια παγκόσμια προσπάθεια, που έχει ως σκοπό την ανάλυση της δομής του ανθρώπινου DNA και τον προσδιορισμό της θέσης των γονιδίων του ανθρώπου. Με την ονομασία human genome project (μελέτη του ανθρώπινου γονιδιώματος), η προσπάθεια αυτή ξεκίνησε το 1990. Πρόκειται για μια 15ετή προσπάθεια με σκοπό να ταυτοποιήσει τα 80.000 γονίδια του ανθρώπινου γονιδιώματος και να προσδιορίσει την αλληλουχία των 3 δισεκατομμυρίων βάσεων που συνθέτουν το ανθρώπινο DNA. Η χαρτογράφηση του υλικού αυτού περιλαμβάνει το διαχωρισμό των χρωμοσωμάτων σε μικρότερα τμήματα και την τοποθέτησή τους σε θέσεις που αντιστοιχούν σε συγκεκριμένες λειτουργίες. Μέχρι τώρα έχουν συγκροτηθεί αρκετοί γενετικοί και φυσικοί χάρτες. Μέχρι τα τέλη του 1997 είχε χαρτογραφηθεί η θέση για περίπου 2.300 γονίδια και είχε ολοκληρωθεί η ανεύρεση της αλληλουχίας του 2,5% του ανθρώπινου γονιδιώματος. Ο τελικός σκοπός είναι η ανεύρεση όλων των γονιδίων του DNA και η ανάπτυξη μεθόδων για χρησιμοποίηση των πληροφοριών αυτών στη μελέτη της βιολογίας και ιατρικής του ανθρώπου. Υπολογίζεται ότι το 2003 θα έχει ολοκληρωθεί η ανεύρεση της αλληλουχίας των βάσεων ολόκληρου του ανθρώπινου γονιδιώματος.*

Key words Μελέτη ανθρώπινου γονιδιώματος, γενετικός χάρτης, φυσικός χάρτης, αλληλουχία βάσεων.

1. Introduction

We may never know when people first recognized the existence of heredity. However, a variety of archeological evidence (e.g. primitive art, preserved bones and dried seeds) have provided many insights. The scripts of the Hippocratic school (fifth to fourth century BC) argue that the male semen is formed by numerous parts of the body and is transported through blood vessels to the testicles. Active "humors", as the bearer of hereditary traits, are drawn from various parts of the body. These humors could be healthy or diseased, could be altered in

Corresponding author: O.A. Giotakos, 2 Erifilis street, GR-116 34 Athens, Greece

Αλληλογραφία: Ο.Α. Γιωτάκος, Εριφύλης 2, 116 34 Παγκράτι, Αθήνα

individuals and, in this new form, could be passed on to offspring. Aristotle (384—322 BC) proposed that male semen was formed from blood rather than from each organ and that each generative power resided in a "vital heat" that it contained. As we know, their thinking was not so different from that of Charles Darwin in his formal proposal of the "theory of pangenesis" (1859). More than a century ago, Mendel established the principles of transmission genetics describing the inheritance of phenotypic expression (1866). He showed that unit factors (now called alleles) exist in pairs, exhibit dominant or recessive relationship in determining the expression of traits and must segregate during gamete formation, in such a way that each member of the fillial generation receives only one of the two factors with equal probability. In 1944, however, direct experimental evidence emerged that the nucleic acid DNA serves as the informational basis for the process of heredity. In 1953, J. Watson and F. Crick proposed that the structure of DNA is in the form of a double helix. In 1971, a paper published by H. Smith, D. Nathans and W. Arber marked the beginning of the recombinant DNA era. The paper described the isolation of an enzyme from a strain of bacteria and the use of this enzyme to cleave viral DNA.

During the last years, a worldwide research effort has the goal of analysing the structure of human DNA and determining the location of human genes. In the United States, the consideration of a genome project began in 1986, and in 1988 the National Institutes of Health (NIH) and the US Department of Energy (DOE) created a joint committee to develop, at first, a five-year plan. Called the human genome project, it got under way in 1990. It is a 15-year effort to identify all the estimated 80,000 genes of the human genome and determine the sequence of the 3 billion chemical bases that make up the human DNA, store this information in databases and develop tools for data analysis. Other countries, notably France, Britain and Japan began similar projects, which are coordinated by an international organization, the human genome organization (HUGO). The task is proceeding in a series of well defined stages. The first stage involves the construction of high-resolution genetic maps for each of the 22 autosomes and the sex chromosomes using identified genes, restriction fragment length polymorphisms (RFLPs), sequence tags sites (STSs) and expressed sequence tags (ESTs). A genetic map incorporating about 15,000 markers was completed in 1995. This genetic map, with an average distance of 2 million base pairs of DNA between markers, is being used to organize the contigs generated by physical mapping. The second level is the construction of physical maps, using either the top-down approach, beginning with chromosomes isolated in somatic cell hybrids, or the bottom-up approach of assembling contigs from DNA segments randomly cloned in cosmids or YACs. The goal was the creation of a physical map consisting of 30,000 STSs with intervals of about 100 kb. The ultimate goal of the project is sequencing the 3.2 billion nucleotides of the human genome, a task that may take a decade or more and requires the development of new technology for sequencing DNA, as well as for information storage, analysis, and retrieval.[1—4]

2. Genetic mapping

Genetic maps[5—7] form the essential backbone needed to guide a physical mapping effort. They are constructed by determining how frequently two "markers", such as a physical trait, a particular medical syndrome or a detectable DNA sequence are inherited together. Genes that lie closer together on a chromosome have a much higher chance of being inherited together, than genes that lie further apart. Genetic studies of families, in order to determine how frequently two traits are inherited together, lead to the production of "genetic maps" in which the distance between genes is measured in centimorgans after the American geneticist T.H. Morgan. As a rough guide, 1% recombination (recombination fraction between two loci depends on how far apart they are in physical terms along the DNA molecule) is equivalent to genetic distance of 1 centimorgan and a physical distance of 1 million base pairs (1 megabase). Linkage groups of markers can be created and the distances between them in terms of recombination determined.

Three commonly used types of DNA markers are: restriction fragment length polymorphisms (RFLPs), variable number of tandem repeats (VNTRs) and microsatellite polymorphisms based on di-, tri-, or tetra-nucleotide repeats. Polymorphisms are variations of the DNA sequence that occur on average once every 300—500 bp. Most variations occur within introns and have little or no effect on an organism's appearance or function, yet they are detectable at the DNA level and can be used as markers. RFLPs were developed first and consist of the presence or absence of a restriction site for a bacterial restriction endonuclease. This is an enzyme that breaks strands of DNA wherever they contain a specific sequence of 6—8 nucleotides. Different enzymes recognize different restriction sites. The locus of interest can be "probed" using a radiolabelled piece of DNA with the same sequence as part of the test locus. This will selectively hybridize to the restriction fragment derived from the locus under examination. The whole process consists of: extracting DNA from white blood cells, digesting the DNA with a restriction enzyme, using gel electrophoresis to separate the fragments by size, denaturating the DNA so that the two strands of each fragment separate, "blotting" the single-stranded DNA onto a filter to immobilize it, probing with a radioactive probe for the test locus which hybridizes to the fragments derived from the test locus that are bound to the filter, washing off unhybridized probe and exposing the filter to a photographic plate which detects the positions of the fragments de-

rived from the test locus as "bands". This process is termed southern blotting, after Dr. Southern. If a restriction site adjacent to the test locus is sometimes present and sometimes absent, then the fragment in which the test locus is found will vary in size and form a polymorphic genetic marker. Whether or not a restriction site is recognized by an enzyme may depend on a single base pair change in the DNA sequence. These restrictions site polymorphisms occur at many sites compared to the limited number of classical genetic markers that were available. They often occur in non-coding regions ("junk" DNA), so that the variation has no biological effect, but acts only as a marker.

Another DNA polymorphism, that can be used as a genetic marker, is the variable number of tandem repeats (VNTRs). In non coding regions there may be certain sequences of DNA which are repeated many times and the number of times may vary. This will produce variation in the size of the restriction fragment containing these repeats, which again may be detected by southern blotting.

A DNA polymorphism that has become increasingly popular is the microsatellite repeat. This consists of a variable number of repetitions of a very small number of base pairs (di-, tri- or tetra-nucleotide repeats) often consisting of cytosine and adenosine (CA-repeats). These repeat sequence polymorphisms are detected by the polymerase chain reaction (PCR). PCR[8] can be used to produce large number of copies of a specific small region of DNA containing the test locus. The specificity of the reaction is defined by a pair of oligonucleotide primers, each about twenty bases long, that match the sequence at either end of the region to be amplified (which would contain the microsatellite repeat). Repeated cycles of denaturation, annealing and elongation result in an exponential increase in the number of copies of the region amplified. The new copies can be radiolabelled during the PCR process, and they are usually so small that when they are size-separated, using electrophoresis, the size differences due to variations in the number of repeats contained in the fragments, are readily detectable. These microsatellite polymorphisms are highly informative because: tiny quantities of DNA are used, alleles can be read very reliably, there is a large number of polymorphic loci and there is often a large number of alleles. New generations of maps are using biallelic markers (singe nucleotide substitution). These are less informative but more common, about every 1,000 bp.

It was estimated that 3,000 well-spaced and informative markers will be needed to achieve a completely linked map, with an average of one centimorgan apart. The latest generation human maps contains 5,264 markers at 2,335 positions at an average distance of 0.75 cM. The most useful genetic markers are those with high heterozygosity, i.e., have many alleles so that most individuals carry different alleles. Each marker should be identified by a sequence-tagged site (STS) as defined below. There are two kinds of genetic marker studies that aim to localize genes influencing susceptibility to an illness: (a) in association studies, many un-related affected individuals are studied and association is said to be present if a particular allele is present more often than in general population or in a matched sample of unaffected controls (b) in linkage studies, related individuals are studied, either siblings or extended pedigrees. Linkage is present when the alleles of a marker tend to co-segregate with a disease within the family. If two markers are at loci which are (in physical terms) close together on the same chromosome, then (in genetic terms) they may demonstrate linkage.

A genome map published in 1987 consisted almost entirely of RFLPs (397, plus 5 protein polymorphisms), and only 7% of these markers had heterozygosities of 70% or greater. In 1992, the NIH/CEPH collaborative mapping group[9] constructed a genetic map of the human genome that consisted of 1,416 loci and included 279 genes and expressed sequences. A total of 339 microsatellite repeat markers assayed by PCR were contained within the map and, of the 351 markers with heterozygosities of at least 70%, 205 were microsatellites. In 1996 C. Dib et al[10] reported the last version of the Genethon human linkage map. This map consisted of 5,264 short tandem (AC/TG)n repeat polymorphisms, with a mean heterozygosity of 70%. The map spans a genetic distance of 3,699 cM and comprises 2,335 positions, of which 2,032 could be ordered with an odds ratio of at least 1,000:1 against alternative orders. The average interval size is 1.6 cM, 59% of the map is covered by intervals of 2 cM at most and 1% remains in intervals above 10 cM.

3. Physical mapping

The physical maps are the basis for the isolation and characterization of individual genes or other DNA regions of interest. They are used to provide the starting material for DNA sequencing. The ability to construct physical maps derives from the recombinant DNA techniques that allow the isolation and cloning of DNA fragments, the identification of specific sequence markers on DNA and the determination of the order of the distance between such markers on a chromosome.

The physical maps can be categorized into two general types. First, the cytogenetic map describes the order and spacing of markers on a DNA molecule. Based on microscopic analysis, cytogenetic maps record the location of genes or DNA markers relative to visible landmarks on the chromosomes. This is the oldest type of physical map and the resolution is rather low (10 million base pairs). Improvement with fluorescence in situ hybridization (FISH) methods[11] allows orientation of DNA sequences

that lie as close as 2—5 Mb. Modifications of the *in situ* hybridization methods, using chromosomes at a stage in cell division (interphase) when they are less compact, increase map resolution to around 100,000 bp.

The second type of a physical map consists of a collection of cloned pieces of DNA, that represent a complete chromosome or chromosomal segment including information about the order of the cloned pieces. There are a variety of techniques for cloning DNA and a number of methods for determining the order of the clones. The technology for constructing overlapping clone sets (known as "contigs") is continuously improving. Techniques as the pulse-field gel electrophoresis, the yeast artificial chromosome (YAC) cloning, the polymerase chain reaction (PCR), the fluorescent in situ hybridization (FISH), and the radiation hybrid analysis, have made the construction of physical maps of large genomes significantly easier.[12,13] One technological barrier was the relatively short length of DNA over which a continuous, or uninterrupted, set of overlapping clones can be readily established. Contigs were typically small, consisting of between two and six cosmid clones (cosmid is a type of vector that can carry a maximum of 40,000 bp). To be more useful, the length of DNA, over which the physical map shows continuity or connectivity, must be considerably longer. Technological improvements now make possible the cloning of large DNA pieces, using artificially constructed chromosome vectors that carry human DNA fragments as large as 1 Mb. These vectors are maintained in yeast cells as artificial chromosomes (YACs).

Another difficulty was the inability to compare the results of one mapping method directly with those of another and to combine maps constructed by two different techniques into a single map. According to the proposal system of Olsom et al (1989),[14] data from any of a variety of physical mapping techniques can be reported in a common language. In this system, any mapped element (individual clone, contig, or sequenced region) is defined by a unique sequence tagged site or STS, which is basically a short DNA sequence that has been shown to be unique. The map is then constructed showing the order and spacing of the STSs. The STS system facilitates the integration of results from different laboratories, regardless of the method used, it produces a single and useful physical map, and it establishes a uniform criteria for determining how complete the map of a particular region is. An STS map, with one STS characterized every 100,000 base pairs was an achievable goal. Finally, an STS map is the appropriate starting point for DNA sequencing.

Christine Bellanne-Chantelot et al[15] in 1992 demonstrated that by using large insert yeast artificial chromosomes (YACs), a whole genome approach becomes feasible. 22,000 YACs of 810 kb mean size (5 genome equivalents) have been fingerprinted to obtain individual patterns of restriction fragments. The same year, a continuous array of overlapping clones covering the entire human chromosome 21q was constructed from YAC libraries, using STSs as landmarks specifically detected by PCR.[16] The YAC contig unit started with pericentrometric and ended with subtelomeric loci of 21q. In 1993, D. Cohen et al[17] presented a first generation physical map of the human genome. They analyzed the CEPH's YAC library, which contained 33,000 clones, the insert size of which were individually determined. These YACs had an average length of 0.9 megabases and covered the equivalent of 10 haploid genomes. The library was screened with 2,100 polymorphic genetically mapped STSs. They also used another approach for rapid and extensive single-copy landmark screening, based on the use of YACs as hybridization probes. Finally, about 500 YACs containing genetically mapped polymorphic STSs (one every 7.4 cM) were positioned on metaphase chromosomes using FISH. This allowed the integration of genetic, physical and cytogenic maps. By screening the library with STSs corresponding to genetic loci, scientists obtained at least one YAC clone for most of them. Taking into account the average 0.9 Mb size for YACs and a quasiuniform distribution of genetic markers, this method provides physical coverage of 20—30% of the genome with ordered YAC clones. Chumacof I et al,[13] analyzed a large insert YAC library by three different experimental procedures: (a) STS content mapping, involving PCR-based screening with genetically mapped microsatellite markers. YACs identified as containing such markers were referred to as genetically anchored YACs, (b) cross-hybridization, involving hybridizing the library with probes derived from individual YACs, and (c) fingerprinting, involving characterizing each YAC in terms of a pattern of restriction fragments detected by two human repetitive sequence probes. These three procedures provided different ways of establishing links, representing potential overlaps between clones. It is not possible to construct a physical map based solely on the complete collection of links. Most YACs aggregate into a few huge, branched, artifactual contigs, because of the high rate of YAC chimerism (40—50%), intra- or interchromosomal sequence similarities in the human genome and the possibility of laboratory error. The authors analyzed a YAC library containing 33,000 clones, with an average insert size of 1 Mb. The development of this strategy resulted in an YAC contig map covering about 75% of the human genome with 225 contigs having an average size of about 10 Mb.

In 1995, Hudson T et al[18] constructed a physical map of the human genome containing 15,086 STSs with an average spacing of 199 kb. The project involved the assembly of a radiation hybrid map containing 6,193 loci and incorporated a genetic linkage map containing 5,264 loci. The authors used a three-step procedure: (a) STSs were assembled into dually linked contings, i.e. groups of STSs connected by dually linkage, (b) the

dually linked contigs were localized within the genome on the basis of a radiation hybrid and genetic map information about loci in the contig, (c) single linkage was then used to join contigs localized to the same small genomic region. By this technique, physical maps of human chromosomes have been reported. Asheorth et al[19] in 1995, generated a metric physical map of human chromosome 19 with overlapping cosmids (contigs), generated by automated fingerprinting, spanning over 95% of the euchromatin, of about 50 Mb. Distances between selected cosmid clones were estimated using FISH in sperm pronuclei. Various types of larger insert clones were used to span gaps between contigs. Over 450 genes, genetic markers, STSs and other markers had been positioned. The same year, Gemmill et al[20] reported the construction of a map of human chromosome 3 which contains both physical and genetic data. The map consisted of 972 megabase-sized YACs, identified with 593 primary markers, of which 162 were highly polymorphic STSs. The remaining markers were hybridization-based. Chromosome 3 was represented by 24 large YAC contigs, and the hybridization- and STS-based datasets covered about 80% (over 160 Mb) of the chromosome. The physical mapping goal is to establish a marker every 100,000 bases across each chromosome (about 30,000 markers). The most complete map yet was published in the summer of 1997 and featured about 8,000 landmarks, that provided twice the resolution of previous maps.[21]

4. DNA sequencing

The ultimate physical map of the human genome is the complete DNA sequence i.e., the determination of all base pairs on each chromosome. Human chromosomes cannot be sequenced directly. Rather, human DNA must be extracted, randomly fragmented and cloned into vectors, capable of stable propagation in a suitable host, such as the bacterium *E. coli* or yeast. Before sequencing, clones must be selected from libraries with chromosomal markers as probes, verified for their fidelity to the genome and designed in a minimal-overlapping way, spanning a portion of a chromosome.[22–24]

The ideal clone library for genomic sequencing has the following features:[25] (a) the clones are highly redundant, covering the human genome many times, (b) the clone coverage is random, and (c) the clones are stable, not subject to deletion or rearrangement during the propagation process. Beginning with a source clone, most large scale sequencing centers use the following steps for sequence determination: randomly fragmenting the source clone into small pieces (1,500 bp), subcloning the small pieces into a sequence-ready vector, sequencing 10—30 subclones per kb of the source clone, and assembling the overlapping sequence data into a contiguous multiple sequence alignment from which a consensus sequence can be inferred from the highest quality data.

Since coding sequences of genes represent most of the potentially useful information content of the genome, but, are only a fraction of the total DNA, some investigators have begun partial sequencing of cDNAs instead of random genomic DNA. cDNAs are derived from mRNA sequences, which are the transcription products of genes. In addition to providing unique markers, these partial sequences, termed expressed sequence tags (ESTs), also identify expressed genes. Other applications of the EST approach include determining locations of genes along chromosomes and identifying coding regions in genomic sequences. Closing gaps, may be the most difficult challenge for sequencing the human genome, as a result of nonrandomness of the clone libraries and STS maps. "Chromosome walking," one strategy for filling gaps, involves hybridizing a primer of known sequence to a clone from an unordered genomic library and synthesizing a short complementary strand. The complementary strand is then sequenced and its end used as the next primer for further "walking". Chromosome walking is also used to locate specific genes by sequencing the chromosomal segments between markers that flank the gene of interest. In 1996, it was proposed that an extensive up-front characterization of a highly redundant BAC clone library would provide a simple and easily automatable approach to the construction of minimal tiling paths. With a library covering the genome 15-fold (300,000 clones), a BAC-end sequence (sequence-tagged connector, or STC) would be found in the genome 5kb every. The STCs are ideal potential chromosomal markers for creating a more dense physical map. 30,000 random STSs, for example, could be localized to a human chromosome by radiation hybrid mapping at an average spacing of 100 kb. The expressed genes from hundreds of different human tissues have been partially sequenced after coping the messenger RNAs into complementary DNA libraries. About 800,000 of these, so called expressed sequence tags (ESTs), are available in public databases. These represent perhaps 40,000—50,000 genes of the estimated total of 70,000—100,000 human genes. ESTs from a variety of model organisms are also available.[25]

In an effort to identify the new genes and analyze their expression patterns, Adams et al,[26] generated 174,472 partial cDNA sequenses, from cDNA libraries that were constructed out of 37 distinct organs and tissues, (ESTs) totaling more than 52 million nucleotides of human DNA sequence. 30 tissues were sampled by over 1,000 ESTs each. Only 8 genes were matched by ESTs from all 30 tissues, and 227 genes were represented in 29 or more of the tissues sampled with more than 1,000 ESTs. About 40% of the identified human genes appear to be associated with basic energy metabolism, cell structure, homeostasis and cell division, 22% with RNA and pro-

tein synthesis, and 12% with cell signaling and communication. Single pass partial sequencing of cDNA clones, from one or both ends, to generate ESTs provides a rapid method of gene discovery that has been widely applied in humans and other species. Messenger RNA species are present at different concentrations in cells, and these differences are reflected in the composition of cDNA libraries. Thus, random-sampling strategies result in abundant mRNAs represented by many ESTs. The authors treated the ESTs as shotgun fragments. cDNA contiguous, overlapping sets of DNA clones (contigs), were assembled based on stringent overlap criteria. These contigs have been termed tentative human consensus sequences (THCs).

5. Future perspectives

At the beginning of 1998 scientists were half way through the time for completing the human genome project. Significant progress has been made, particularly in identifying and mapping genes, developing a stable DNA sequencing technology, and in building the computational tools required for the analysis of sequence data.[27] The large-scale sequencing of the 3.2 billion base pairs of the human genome has recently begun. Approximately 60 million base pairs had been analyzed, and roughly 2.5% of the human genome had been sequenced until 1998. The genomes of E. coli, yeast, and 11 microbes have been completely sequenced. Those of the worm, fruit fly, mouse, and human have been partially sequenced. These sequences have dramatically altered the practice of genetics, molecular biology, developmental biology, immunology, and microbiology. To complete the genome by 2005, starting in 1998, seven large-scale sequencing centers, for example, would each have to complete sequences on the order of 75 Mb/year. Sequencing centers now have a throughput of 2—30 Mb/year. If the genome is to be sequenced on time and within budget, sequencing must become significantly faster and cheaper. The cost of routine large-scale sequencing will ultimately have to be reduced to well below 50 cents per base pair.

The current standards for sequencing, set by the NIH and DOE: require that three conditions must be met,[25] (a) error rate of no more than 1 in 10,000, (b) sequence contiguity, that is sequence without gaps, and (c) clone validation, that is, a demonstration that clones faithfully represent the genome. Currently, established sequences are available from four large databases, two in USA, one in Europe and one in Japan. In addition to new insights into human biology, the genome project has focused its attention on ethical, legal and social consequences of genomic research. These issues include the application of genetic testing, privacy and the fair use of genetic information in insurance, employment and health care. The ELSI (ethical, legal and social implications) program of the human genome project has been set up to address these issues.

According to the "5-year research goals 1998—2003, US human genome project",[21] it is estimating to:

a. Read one-third of the human DNA sequence by the end of 2001, and the complete human genome sequence by the end of 2003. Generate sets of full-length cDNA clones and sequences that represent human genes and model organisms.

b. Complete the sequence of the roundworm *C. elegans* genome by 1998, the sequence of the fruitfly Drosophila genome by 2002, and the sequence of the mouse genome by 2008.

c. Identify common variants in the coding regions of known genes. Although the vast majority of our DNA is identical a mong individuals sequence variations can have a major impact on how our bodies respond to diseases, environmental insults, such as bacteria and viral infections, and drugs and other therapies.

d. Examine issues raised by novel genetic technology and information into health care and public health activities. Explore how new genetic knowledge may interact with a variety of philosophical, theological, and ethical perspectives.

If successful, the completion of the human DNA sequence in 2003 will coincide with the 50th anniversary of Watson and Crick's description of the fundamental structure of DNA. The complete genome sequence will provide useful data to the biological and medical communities. Indeed, it must be seen as the beggining of a new era in biological research and not as the end of it.[28—30]

Glossary

allele: one of the possible mutational states of a gene, distinguished from the other alleles by phenotypic effects.

bacterial artificial chromosome (BAC): A vector used to clone DNA fragments (100—300 kb insert size) in *Escherichia coli* cells.

base pair (bp): two nitrogenous bases (adenine and thymine or guanine and cytocine) held together by weak bonds.

cDNA: DNA synthesized from an RNA template by the enzyme reverse transcriptase.

centimorgan (cM): a unit of distance between genes on chromosomes. One centimorgan represents a value of 1% crossing-over between genes.

chromosome map: a diagram showing the location of genes on chromosomes.

cloned library: a collection of cloned DNA molecules representing all or part of an indivindual's genome.

contig: group of cloned pieces of DNA representing overlapping regions of a particular chromosome.

cosmid: a vector designed to allow cloning of large segments of foreign DNA. Cosmids are hybrids composed of the *cos*-sites of lambda phage inserted into a plasmid.

denaturated DNA: DNA molecules that have been separated into single strands.

DNA polymerase: an enzyme that catalyzes the synthesis of DNA from deoxyribonucleotides and a template DNA molecule.

DNA sequence: the relative order of base pairs, whe-ther in a fragment of DNA, a gene, a chromosome, or an entire genome.

exon: the DNA segment of a gene that is transcribed and translated into protein.

expressed sequence tags (EST): are STSs derived from cDNAs.

fluorescence in situ hybridization (FISH): a physical mapping approach that uses fluorescein to detect hybridization of probes with metaphase chromosomes and with the less condensed somatic interphase chromatin.

gene: the fundamental physical and functional unit of heredity.

gene code: the sequence of nucleotides, coded in triplets (codons) along the mRNA, that determines the sequence of amino acids in protein synthesis.

genome: the array of genes carried by an individual cell.

kilobase (kb): A unit of length consisting of 1.000 nucleotides.

library: An unordered collection of clones, whose relationship to each other can be established by physical mapping.

linkage: condition in which two or more nonallelic genes tend to be inherited together.

linkage map: a map of the relative positions of genetic loci on a chromosome, determined on the basis of how often the loci are inherited together.

locus: the site or place on a chromosome where a particular gene is located.

megabase (Mb): unit of length for DNA fragments equal to 1 million nucleotides and roughly equal to 1 cM.

plasmid: autonomously existing, and replicating, extrachromosomal circular DNA molecules, distinct from the normal bacterial genome and nonessential for cell survival under nonselective conditions.

polymerase chain reaction (PCR): a method of amplifying DNA segments that uses cycles of denaturation, annealing, and DNA polymerase-directed DNA synthesis.

restriction fragment length polymorphism (RFLP): variation between individuals in DNA fragment sizes cut by specific restriction enzymes. Polymorphic sequences that result in RFLPs are used as markers on both physical maps and genetic linkage maps.

sequence tagged site (STS): short (200—500 bp) DNA sequences that have a single occurence in the human genome and whose location and base sequence are known.

sequencing: determination of the order of nucleotides (base sequence) in DNA, RNA or the order of aminoacids in proteins.

tanscription: the synthesis of an RNA copy from a sequence of DNA (a gene). The first step in gene expression.

translation: the process in which the genetic code curried by mRNA directs the synthesis of proteins from amino acids.

vector: in recombinant DNA technology, an agent such as a phage or plasmid into which a foreign DNA segment can be inserted.

yeast artificial chromosome (YAC): a vector used to clone DNA fragments (up to 400 kb). It is constructed from the telomeric, centromeric and replication origin sequences needed for replication in yeast cells.

Βιβλιοχραφία

1. Clung WS, Cummings MR. "An introduction to Genetics", "Modification of Mendelian Ratios", Linkage and Chromosome Mapping", and "Recombinant DNA". In: Clung WS, Cummings MR (eds) *Essential of Genetics.* Prentice-Hall, N. Jersey, 1996:2—17, 68—89, 112—141, 362—385
2. Understanding our genetic inheritance. The Human Genome Project: the first five years 1991—1995. NIH Publ, 1990: 5—21
3. Collins F, Galas D. A five-year plan for the US human genome project. *Science* 1993, 262:43—46
4. Read PA, Brown T. "Gene analysis techniques", Cloning in higher organisms" and Future prospects". In: Williams J, Ceccarelli A, Spurr N (eds) *Genetic Engineering.* Biol Sc Publ, Oxford, 1996:43—60, 95—110, 111—119
5. Francke U, Lalouel JM, With R. Gene mapping, Analysis of genetic linkage. In: Emery A, Rimoin D, Livingstone Ch

(eds) *Principles and practice of Medical Genetics.* UK, 1992:133—148, 149—164
6. Corey P. Chromosome Mapping. In: Radford A, Cove D, Baumberg SA (eds) *Primer of Genetics.* Longman UK, 1995:23—48
7. Ott J. Strategies for characterizing highly polymorphic markers in human genome mapping. *Am Hum Genet* 1992, 51: 283—290
8. Erlich HA, Gerald D, Sninsky J. Recent advances in the polymerase chain reaction. *Science* 1991, 252:1643—1651
9. NIH/CEPH Collaborative Mapping Group: A comprehensive genetic linkage map of the human genome. *Science* 1992, 258:67—76
10. Dib C, Faure S, Fizames C et al. A comprehensive genetic map of the human genome based on 5,264 microsatellites. *Nature* 1996, 380:152—154
11. Le Beau M. One FISH, two FISH, red FISH, blue FISH. *Nature Genetics* 1996, 12:341—344
12. Adams M. Complementary DNA sequencing: EST and human genome project. *Science* 1991, 252:1651—1656
13. Chumacov IM, Rigault P, Le Gall I et al. A YAC conting map of the human genome. *Nature* 1995, 377:175—186
14. Olson R. The STS system. *Science* 1989, 254:1434—1435
15. Bellanne-Chantelot C. Mapping the whole human genome by fingerprinting YACs. *Cell* 1992, 70:1059—1068
16. Chumacov I. Continum of overlapping clones spanning the entire human chromosome 21q. *Nature* 1992, 359:380—386
17. Cohen D. A first generation physical map of the human genome. *Nature* 1993, 366:698—701
18. Hudson T, Stein L, Gerety S et al. An STS-Based map of the human genome. *Science* 1995, 270:1945—1954
19. Ashworth L. An intergrated metric physical map of human chromosome 19. *Nature Genetics* 1995, 11:422—427
20. Gemmill R. A second-generation YAC contig map of human chromosome 3. *Nature* 1995, 377:299—302
21. http://www.ornl.gov/TechResources/Human_Genome/
22. Jordan E, Collins F. A march of genetic maps. *Nature* 1996, 380:111—112
23. Thomas S, Davies A, Birtwistle N et al. Ownership of the human genome. *Nature* 1996, 380:387—388
24. Oliver S. From DNA sequence to biological function. *Science* 1989, 254:1434—1435
25. Rowen L, Mahairas G, Hood L. Sequencing the human genome. *Science* 1997, 278:605—607
26. Adams M. Initial assessment of human gene diversity and expression patterns based upon 83 million nucleotides of cDNA sequence. *Nature* 1995, 377:3—16
27. Hawkins T. A magnetic attraction to high-throughput genomics. *Science* 1997, 276:1887—1889
28. Boguski M, Schuler G. Establishing a human transcript map. *Nature Genetics* 1995, 10:369—371
29. Gibbs R. Pressing ahead with human genome sequencing. *Nature Genetics* 1995, 11:123—125
30. Cox D, Myers R. A map of the future. *Nature Genetics* 1996, 12:117—118

Υποβλήθηκε 9.4.1998
Εγκρίθηκε 5.3.1999

Η Εταιρεία Ιατρικών Σπουδών και οι Ιατρικές Εκδόσεις ΒΗΤΑ πληροφορούν τους συγγραφείς των άρθρων που δημοσιεύονται στην Ιατρική ότι οι αντίστοιχες αγγλικές περιλήψεις εμφανίζονται και στο Internet: www.mednet.gr/beta/beta.htm