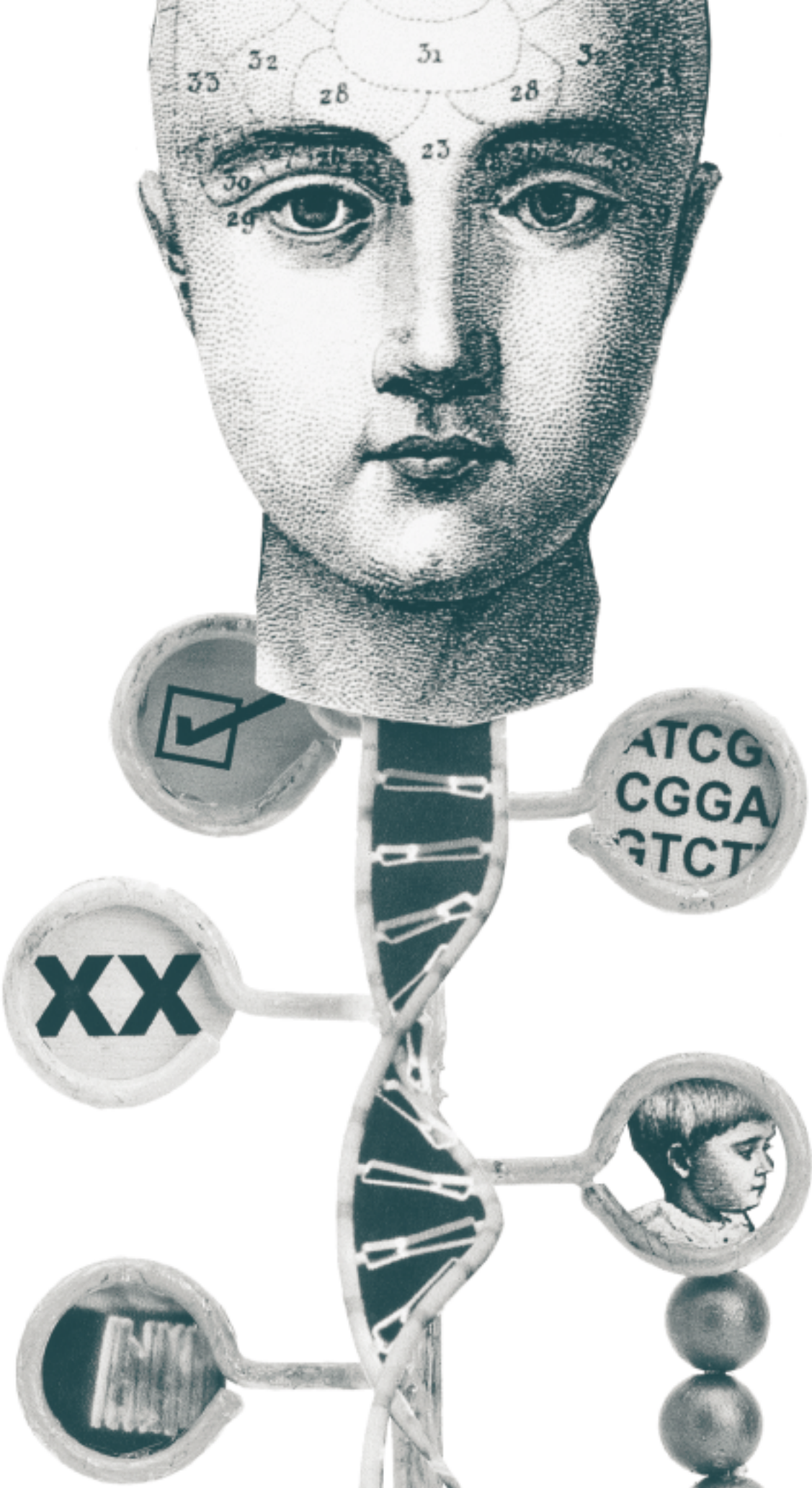

31
32
32
33
28
28
23
30
29
29
ATCG
CGGA
GTCT
XX

The Human Genome Project

BY **STEPHEN SCHERER**

RÉSUMÉ ► Il existe désormais une séquence préliminaire de l'ensemble du génome humain et de la plupart des gènes encodés. Dans le présent article, nous expliquons ce qu'est le génome, ce que représente son séquençage et comment il a été réalisé, avant d'aborder une réflexion sur l'incidence de cette découverte sur notre façon de faire de la recherche scientifique et, sur un plan plus fondamental, sur notre connaissance de nous-mêmes. La description complète du génome humain constitue le fondement de la biologie humaine et la condition essentielle à la connaissance approfondie des mécanismes de la maladie. Comme telle, l'information découlant du séquençage du génome humain équivaudra à un guide de référence pour la science biomédicale XXIe siècle. Elle aidera les chercheurs et les cliniciens à comprendre, diagnostiquer et finalement traiter nombre des 5 000 maladies génétiques qui affligent l'humanité, notamment les maladies plurifactorielles dans lesquelles la prédisposition génétique et les éléments environnementaux jouent un rôle important. (Traduction : **www.isuma.net**)

ABSTRACT ► We now have a draft sequence of the entire human genome and most of the genes it encodes. Here we explain what the genome is, what it means to sequence it, outline how the sequencing was done, and reflect on what this means for the way we do science and, more fundamentally, how we understand ourselves. A comprehensive description of the human genome is the foundation of human biology and the essential prerequisite for an in-depth understanding of disease mechanisms. As such, information generated by the sequenced human genome will represent a source book for biomedical science in the 21st century. It will help scientists and clinicians to understand, diagnose and eventually treat many of the 5000 genetic diseases that afflict humankind, including the multifactorial diseases in which genetic predisposition and environmental cues play an important role.

ILLUSTRATIONS : KATY LEMAY

GENES ARE instructions that give organisms their characteristics. The instructions are stored in each cell of every living organism in a long string-like molecule called Deoxyribonucleic Acid (DNA). DNA molecules are subdivided into finite structures called chromosomes (Figure 1). Each organism has a characteristic number of chromosomes. For humans the number is 46 (23 pairs) and this complete set of genetic information is called the genome (Figure 2).

DNA, genes and genomes

Human DNA looks like a twisted ladder with three billion rungs. If unwound, your DNA would stretch over five feet, but it is only 50 trillionths of an inch wide. The total amount of DNA in the 100 trillion or so cells in the human body laid end to end would run to the sun and back some 20 times. The three billion rungs are made up of chemical units, called "base pairs," of nucleotides — adenines, thymines, cytosines and guanines, represented by the letters A, T, C and G. Particular combinations of these DNA base pairs (or genes) constitute coded instructions for the formation and functioning of proteins, which make up the body and govern its biological functioning (examples of proteins include insulin, collagen, digestive enzymes, etc.). Ribonucleic acid, (RNA) is a single stranded copy of DNA that acts as an intermediate messenger molecule that allows DNA sequence to be translated to protein. This central process, whereby DNA transcribes to RNA, which in turn transcribes to protein, underlies all of life.

Some genes are made up of only a few hundred base pairs, others run to a couple of million base pairs. It is now estimated that we have around 30,000 to 40,000 genes (much more work will be required to determine the precise number), but those genes account for less than five percent of our DNA, with the rest being filler. For lack of a better term, the non-genic DNA is often dubbed "junk" since its role is not fully known (in fact this might be a good term since just like the cell we often keep our "junk" just in case we need it contrary to "garbage" which we, and the cell, throw away).

The Human Genome Project and DNA sequencing

"Sequencing" is the process of determining the specific order and identity of the three billion base pairs in the genome with the ultimate goal of identifying all of the genes. "Mapping" is the process of identifying discrete DNA segments of known position on a chromosome which are then used for sequencing (mapping is a crucial step for proper reconstruction of the genome; it usually precedes sequencing but is also necessary in post-sequencing). To give one a sense of the enormity of the task to decode the human genome, a simple listing of all of the ATCG combinations would fill hundreds of phone books. And having the sequence is only a first step in finding the genes, understanding how they operate, and how particular genetic discrepancies are associated with diseases. Using this DNA sequence information to diagnose, prevent or treat those genetic predispositions or causes of disease is an even more complex task.

When the idea of sequencing the human genome was first proposed about 15 years ago, it was highly controversial. For one thing, given the technology at the time, sequencing DNA was painstakingly slow and expensive. Sequencing the entire human genome seemed impractical. Others felt it made more sense to focus research effort and money only on discovering particular genes thought to be associated with diseases, rather than trying to spell out the entire genome. However, even for disease projects aimed at finding single gene disorders (such as that for cystic fibrosis) the worldwide cost could often run in the tens of millions dollars before success was achieved. The idea for the human genome project eventually won support, justified as an investment in fundamental biological infrastructure upon which the scientific community around the world would build exciting new research that would lead to the diagnosis and treatment of disease.

The enormity of the challenge inspired the creation of an international, publicly funded research initiative called the Human Genome Project (HGP), officially launched in the early 1990s. The Human Genome Organization (HUGO) brings together over 1000 scientists from 50 countries worldwide, and has a mandate that includes promoting international collaboration within the HGP. The earliest genome meetings were attended by a rag-tag group of usually fewer than 100 biologists of diverse background

FIGURE 1

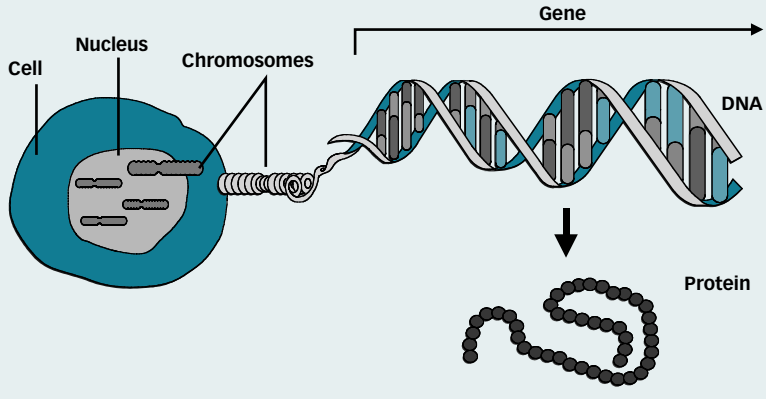


Figure 1. The human genome consists of 23 pairs of chromosomes that are comprised of DNA. Approximately 5% of the DNA along the chromosome encodes for genes. Genes make proteins. Proteins provide cellular structure and functional activity within the body.

FIGURE 2

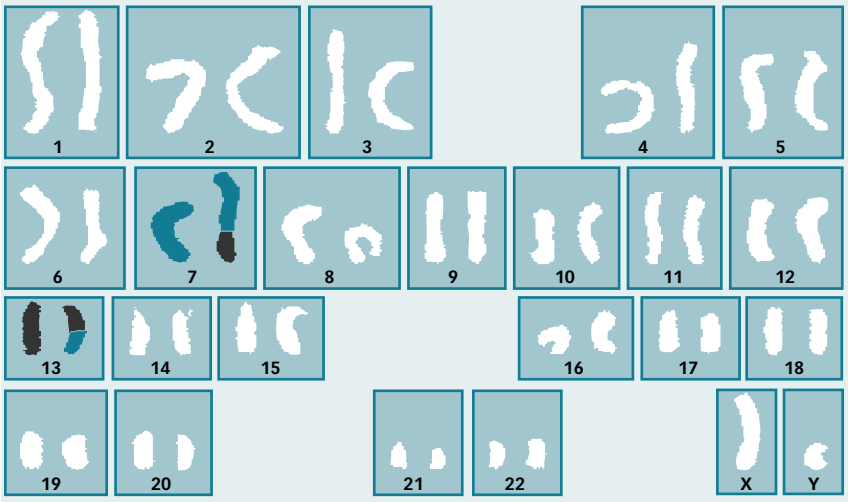


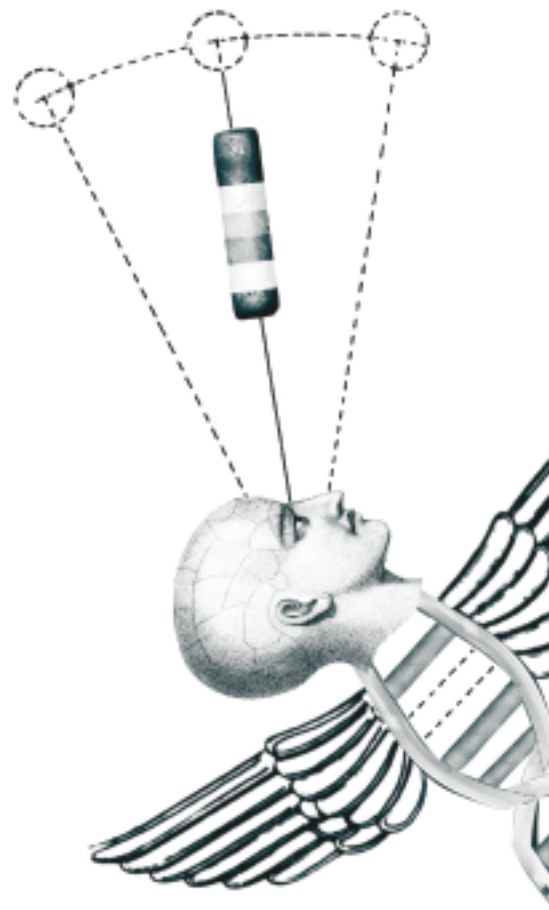
Figure 2. The 23 pairs of human chromosomes one from each pair being inherited from one parent. Note the interchange of material between chromosomes 7 and 13 which disrupts a gene leading to disease.

and research interests, only a few of whom had some type of vision of how to accomplish this distant goal (Figure 3). Since 1990, the strategies and measures of successes of the HGP have most often been debated at an annual meeting held in Cold Spring Harbor Laboratories, New York (directed by James Watson, co-discover of DNA). The most accurate documentation of the science and progress behind the genome project can be found within the abstracts of this meeting. The early consensus was that before large-scale and cost-effective sequencing of the entire genome could be completed, numerous technological developments, associated with mapping and sequencing processes as well as with the computational capacity to make sense of the resulting information, were essential.

With developments in technology over the last few years, the seqencing is now done through a process called fluorescence-based dideoxy sequencing (Figure 4a). Fragments of DNA are first cloned in bacteria. They are then put into a chemical reaction with free nucleotides, some of which are tagged with fluorescent dyes. Nucleotides attach themselves to the DNA fragments in a particular order. Similarly, dyed nucleotides can attach themselves to the DNA fragments, but other nucleotides will not adhere to the dyed nucleotides. Thus the chemical reaction generates DNA fragments of various length that each terminate at fluourescently dyed A, T, C or G nucleotides. The underlying sequence for the range of the DNA fragments created in the chemical reaction is then determined by an automated sequencing machine. The synthesized DNA fragments are negatively charged. The sequencing machine sets up an electrical field. The DNA fragments move through a pourous gel toward the positive charge, with shorter fragments moving more quickly through the gel. The fluorescent, tagged bases at the end of the fragments can be detected with the help of a laser, and the resulting information stored on a computer. As a result, the order in which the particular tagged nucleotides are read reflects their order on the stretch of DNA that has been replicated in the chemical reaction.

Each reaction reveals the sequence of about 500 letters (A,T,C,G) of DNA before the process runs its course. Once these relatively tiny sequences are obtained, their place in the overall genome DNA sequence must be deter-

It will likely take two to three years for the draft form of the DNA sequence to reach what is considered to be in a finished state.



Having the sequence is only a first step in finding the genes, understanding how they operate, and how particular genetic discrepancies are associated with diseases.

FIGURE 3

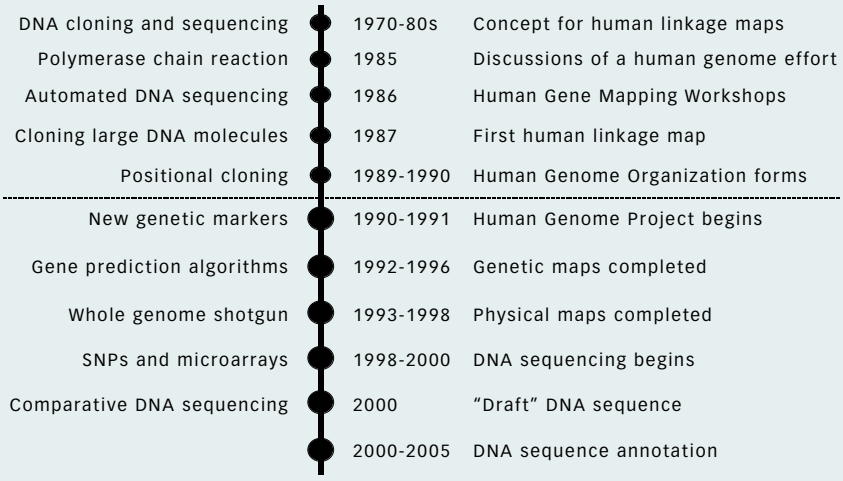


Figure 3. Timeline of the Human Genome Project (HGP). A historical summary of some of the enabling technologies (on the left) and the achievements (on the right) leading up to and including the HGP are shown. The formal international HGP began around 1990. Other aspects of the HGP are the study of model organisms, analysis of genome variation, and the development of bioinformatics. Establishing training and public education programs, as well as the study of the ethical, legal, and social issues of genetics research in society are also important priorities.

mined. To achieve a working draft DNA sequence of the genome, two approaches were followed. The HGP began by creating a detailed map that would provide a framework for the subsequent DNA sequencing. Numerous DNA "markers," that is segments with an identifiable location on the chromosome (Figure 4b), were generated through many worldwide initiatives.

This enabled the HGP to break down the genome into about 30,000 sections, each containing an average of 100,000 to 200,000 base pairs (Figure 5). For the actual sequencing, each of these sections was broken down into still smaller fragments, of about 2000 base pairs. Initially, the plan was to put the fragments to be sequenced in order, followed by complete sequence determination of each fragment in a systematic manner so that the entire human DNA sequence would be revealed. This 'clone-by-clone' method produces highly accurate sequence with few gaps. However, the upfront process of building the sequence maps is costly, time consuming and therefore can be progress-limiting.

In 1998, to accelerate progress, a major deviation from this plan was the decision to collect only partial data from each DNA fragment, hence, a working or rough draft (the change of strategy was partly due to the launching of a privately funded company, Celera, which decided to determine the DNA sequence of small pieces of human DNA totally at random using the "whole genome shotgun" (WGS) approach described below). Working draft DNA sequence usually covers 95 percent of the genome (maintaining 99% accuracy) but it is divided into many unordered segments with gaps between. Additional sequencing is required to generate the finished DNA sequence such that there are no gaps or ambiguities, and the final product is greater than 99.99 percent accurate. At the time of writing, close to 90 percent of the human genome sequence has been determined (approximately 70% of this in working draft and 30% in finished form).

A second approach to generating a draft sequence of the human genome is the whole genome shotgun (WGS) associated with Celera Corporation. In WGS, sufficient DNA sequencing is performed so that each nucleotide of DNA in the genome is covered numerous times in fragments of about 500 base pairs. Identifying where those individual fragments fit in the overall DNA is accomplished through the use of powerful computers that analyze the raw data to find overlaps.

The WGS approach was used by scientists at The Institute for Genomic Research to generate the first complete sequence of a self-replicating organism called *Haemophilus influenzae* as well as many other single-cell organisms. The advantage of WGS is that the upfront steps of constructing maps are not needed. For organisms with much larger and more complex genomes, such as the fruitfly *Drosophila melanogaster* and humans, assemblies of sequences are expected to be complicated by the presence of a vast number of repetitive elements (approximately 50% of human DNA is repeats!). Notwithstanding, Celera initiated a WGS project of the Drosophila genome and in doing so, 3.2 million sequence reads were completed (giving a 12.8 times coverage of the 120 Mb (million bases) genome). Based on these data, 115 Mb of DNA sequence were assembled and, although it was quite comprehensive, the genome was still divid-

FIGURE 4

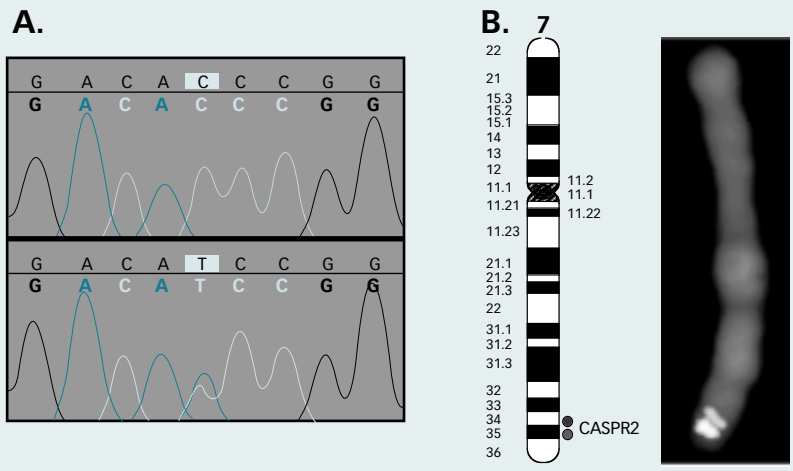


Figure 4. Techniques of the HGP. (A) DNA sequencing and (B) positioning a gene on a chromosome, in this case the CASPR2 gene on human chromosome 7.

If a segment of DNA has been conserved throughout evolution, it almost always encodes a gene or a regulatory element controlling the expression of that gene.

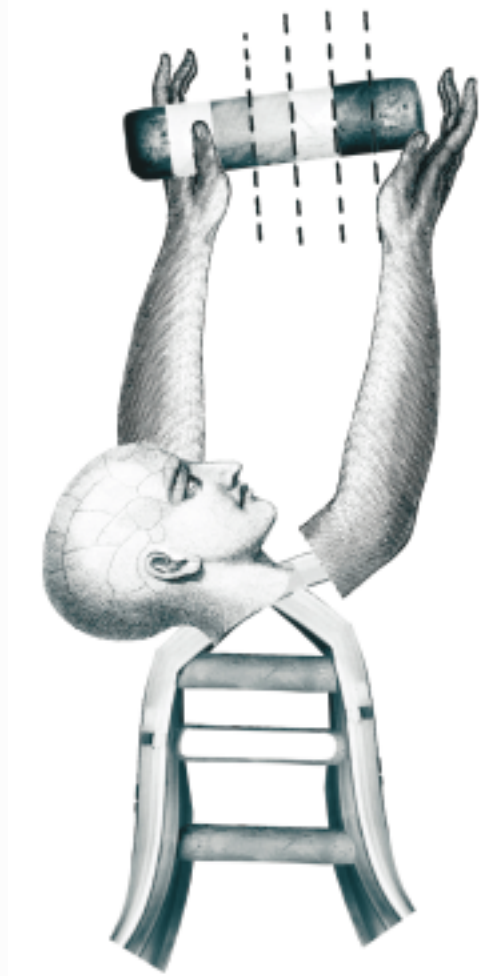
ed by 1630 gaps. Closure of the gaps is being completed with the help of the maps generated by the HGP.

Following the experience gained from the *Drosophila melanogaster* project, Celera planned a whole genome assembly of human DNA. The aim of the project was to produce highly accurate, ordered sequence spanning more than 99.9 percent of the human genome. Based on the size of the human genome and the results from the *Drosophila* experiment, it was predicted that over 70 million sequencing reactions would need to be completed. The alignment of the resulting sequence assemblies along the genome would be accomplished using the large number of DNA markers (including genes, see figures 4b and 5) and physical maps generated by the ongoing HGP. Since sequencing efforts were escalated by the publicly funded HGP (and released each night on the Web) using its approach, Celera could also easily integrate this data into their assemblies thereby greatly reducing the amount of sequencing required. In the end, Celera completed approximately 28 million reads and merged these data with DNA sequence in public databases. The final sequence assemblies were ordered along chromosomes using the DNA markers and maps from the HGP.

It will likely take another two to three years for the draft form of the DNA sequence to reach what is considered to be in a finished state. This work is ongoing around the world in big and small laboratories alike. The completeness and accuracy of these versions of the human genome sequence will be tested by many types of experimentation over the next decade. In future DNA sequencing projects, as is currently ongoing for the mouse genome, a combination of the clone-by-clone and WGS strategies will likely become the standard because they are largely complementary.

Annotation of the human genome DNA sequence

Ascribing functional meaning to all the genes and other important biological information encoded by the DNA in our genome will be the ultimate goal of the HGP. This process is often called "annotation." With this information, it will be possible to investigate the roles of all of the gene products, how they are controlled, how they interact, and how they might be involved in disease. The process of annotating the human DNA sequence



About 5,000 human diseases are known to have a genetic component, and 1,000 disease-associated markers or genes have already been isolated.

FIGURE 5

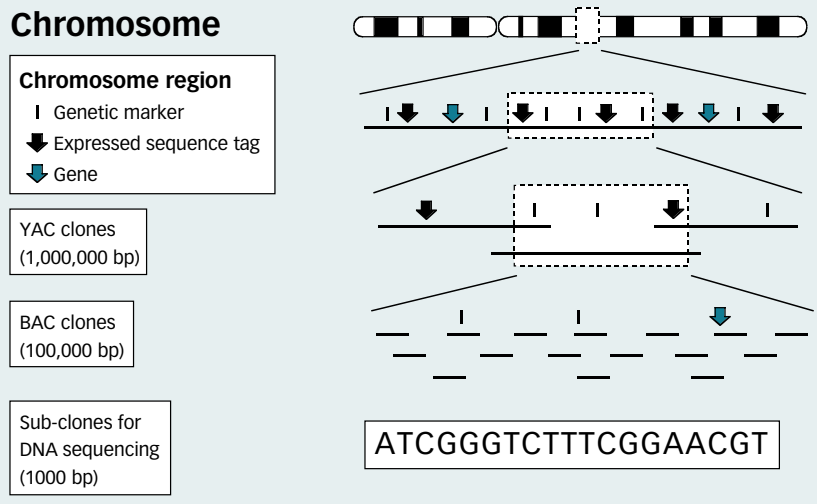


Figure 5. Systematic mapping and sequencing of human chromosomes using a clone-by-clone approach. DNA markers from a chromosome region are positioned and simultaneously used to order large cloned DNA molecules (called YACs and BACs). BACs are broken down in 'sub-clones' for DNA sequencing. Obtaining partial sequence of the BACs represents 'draft' sequence. In the whole genome shotgun approach (WGS) the intermediate steps of cloning and mapping BACs and YACs are not completed. It is estimated there are 3 billion A,T,C,and G's in the human genome.

will take several forms including the catalogue of the genes; the identification of the genes and DNA sequence variations that either directly cause or are associated with disease; and the study of variation between individuals in the population.

Cataloguing the genes: Just as difficult as constructing maps or determining the DNA sequence of the human genome will be identifying all of the genes it encodes. As mentioned earlier, the genes comprise less than five percent of the DNA scattered throughout the three billion nucleotides of genetic information. To complicate matters, human genes almost always appear as discontinuous segments of DNA along a chromosome divided into gene-coding regions (exons) and non-coding regions (introns) (Figure 6). To complicate matters further, there are DNA stretches that do not encode for proteins but, instead, the RNA molecule performs a biological function within the cell. Notwithstanding, most genes have characteristic features that can be identified using computer programs trained to search for these identifiers. For example, genes usually start with the sequence ATG, end with TAG, TGA or TAA, and the boundaries between introns and exons are often defined by the dinucleotides AT and CG, respectively (Figure 6).

Another powerful approach to finding genes is to compare the DNA sequence between two organisms to search for shared elements (Figure 7). If a segment of DNA has been conserved throughout evolution it almost always encodes a gene or a regulatory element controlling the expression of that gene. Almost every human gene known can also be found in the genomes of other mammalian species (e.g., mice and dogs) with the similarity of the two sequences often exceeding 85 percent. The genomes of humans and chimpanzees are 98.5 percent identical. Some genes, including those that encode proteins involved in replicating DNA (and are thus absolutely crucial to survival of every organism), are highly conserved in every species from human all the way down the evolutionary tree to yeast and bacteria. Other genes that originated from a common progenitor have evolved between species to have unique functions such as specific fertilization (Figure 7).

Current estimates put the number of genes in the human genome at between 30,000 and 40,000. The precise number continues to be hotly debated, and it could very well be that there are upwards of 50,000 to 60,000 or more genes in the genome. Discrepancy in numbers can be attributed to different interpretations in the definitions, and the use of different datasets which, in every case, are still largely incomplete. For example, at present we only know the complete sequence from start to end of about 25 percent of the 30,000 to 40,000 genes that have been predicted using computer algorithms.

One of the most interesting observations to come out of genome sequencing is that the number of genes does not necessarily correlate with organism complexity or its place in the evolutionary hierarchy. For example, the fruitfly *Drosophila melanogaster* has 13,000 genes and comprises 10 times more cells than the worm *C. elegans*, which has 19,000 genes.

Disease gene identification: The immediate medical application of

FIGURE 6

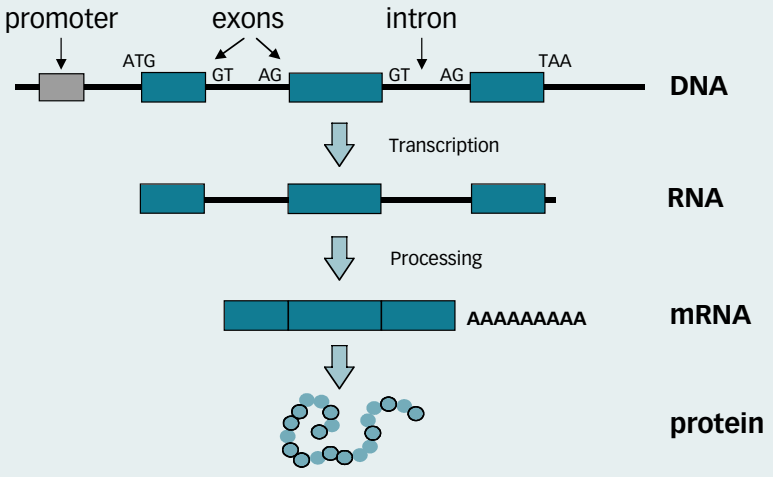


Figure 6. Anatomy of a gene. While genes occupy less than 5% of the DNA on a chromosome they often have characteristic features that allow them to be identified amongst the so-called 'junk' DNA. Promoters are the regions of DNA that turn genes on and off. Most genes start with ATG and end with TAA, TGA, or TAG. The GT and AG di-nucleotides usually flank the parts of genes (exons) that code for proteins. RNA and messenger RNA (mRNA) are intermediates to the production of proteins.

FIGURE 7

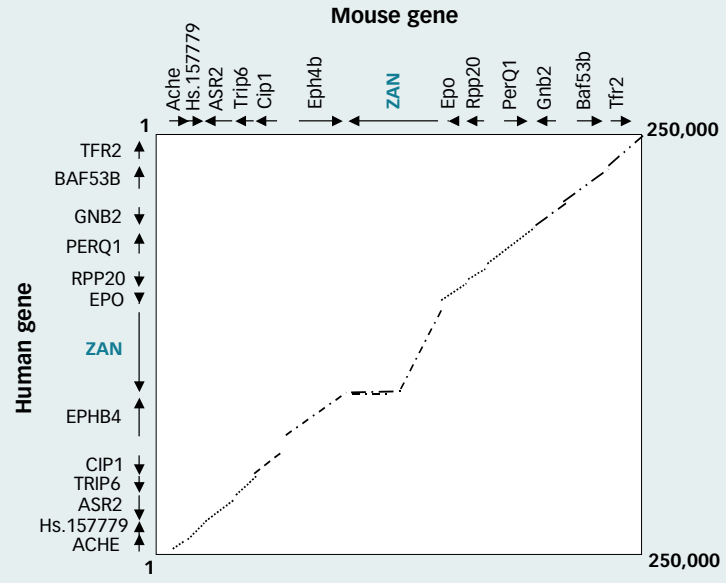


Figure 7. Genes conserved between species. Thirteen human genes on human chromosome 7 are compared to the equivalent genes in the mouse genome. A 45 degree line between indicates the DNA sequence of the human and equivalent mouse gene is >85% identical. The deviation at the intersection of the zonadhesin (ZAN) genes occurs because each encodes a protein that is involved in the process of species-specific fertilization of sperm to egg. Therefore, during evolution the DNA sequences of these genes diverged from each other to fulfill the role of making a protein unique to the survival of that organism.

human genome information is in identification of genes associated with disease, and the pursuit of new diagnostics and treatments for these diseases (Figure 8). About 5,000 human diseases are known to have a genetic component, and 1,000 disease-associated markers or genes have already been isolated over the years. Until the 1980s, the primary strategy for identifying human genetic disease genes was to focus on biochemical and physiological differences between normal and affected individuals. However, for the vast majority of inherited single-gene disorders in humans, biochemical information was either insufficient or too complicated to give insight into the basic biological defect.

In the early 1980s an alternative strategy was introduced that suggested disease genes could be isolated solely on their location along the chromosome. This approach, now called "positional cloning," consists of first determining which of the 23 chromosomes the disease gene resides on. Then genes on the chromosome are systematically tested for DNA sequence changes (or mutations) that occur only in the family members having the disease but which are not found in unaffected individuals.

Prior to the initiation of the HGP, the disease gene for Duchenne/Becker muscular dystrophy, cystic fibrosis and a few others were identified. With the development of new mapping resources and technologies generated by the HGP, the positional cloning process was greatly simplified. This has accelerated the pace of discovery of new disease genes including those for fragile X syndrome, Huntington disease, inherited breast cancer, early onset Alzheimer disease and many others.

With the HGP sequence now well advanced, it becomes possible to dissect the molecular genetics of multifactorial diseases such as cancer and cardiovascular disease which involve multiple combinations of genes and strong environmental components. The study of these common diseases will require careful patient and family data collection, thorough clinical examination, systematic DNA analysis and complex statistical modeling.

We now face the more daunting responsibility of having power over the genetic destiny of our own species.

FIGURE 8

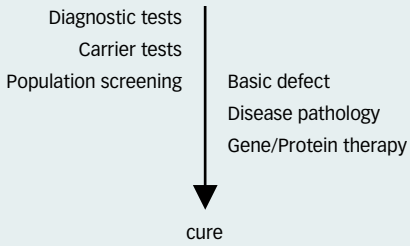


Figure 8. Applications of the Human Genome Project. The immediate application of the human genome information are identification of genes associated with disease, and development of diagnostic and predictive tests. Based on this information the pursuit of new treatments for these diseases can be pursued.

Human genome sequence variation: There are many types of variation in the human genome with single nucleotide changes (SNPs) being the most frequent occurring in one of 1000 nucleotides. Therefore, on average, each genome contains approximately three million SNPs. Other DNA sequence variations include copy number changes, insertions, deletions, duplications, translocations, inversions and more complex rearrangements, but these are not as frequent. There are, thus, significant differences between the genome of the two parents whose chromosomes come together to form new life. This genetic variation is the basis of evolution. It is this genetic variation that contributes, in part, to health and to each individual's unique traits. However, when DNA sequence variants occur within genes or affect their expression, disease can sometimes result.

Technological advances, availability of genetic markers and higher resolution SNP maps generated by the HGP have revolutionized the study of human genetic variation. The biomedical applications include identifying variations within specific genes that cause or predispose to disease, finding gene-environment interactions that might have toxologic implications and identifying variations in immune response genes that will have implications for transplantation and vaccine development.

The study of human genome variability has also greatly improved our grasp of human history, evolution of the human gene pool and our basic understanding of genome evolution. For example, studies of DNA variation have revealed that the vast majority of genetic diversity (>80%) occurs between individuals in the same population even in small or geographically isolated groups. Moreover, most variation seems to predate the time when humans migrated out of Africa around 200,000 years ago. Therefore, the concept of homogeneous groups (or races) having major biological differences is not consistent with genetic evidence.

The Human Genome Project and society

With the success of the HGP, we have overcome the psychological barrier of cracking nature's code and now face the more daunting responsibility of having power over the genetic destiny of our own species. As such, the HGP joins the ranks of the other massive scientific undertakings of the 20th century — splitting of the atom and the conquest of space — in transforming civilization. And, just as Galileo's work was foundational to proving the Copernican theory which debunked the notion that the earth was the centre of the universe, the HGP proves there are human-to-animal DNA sequence links, thus substantiating Darwin's theory that we are not a unique life form. With this information, the HGP promises to give us profound knowledge as the basis to understanding how our minds work,

to be able to quantitate nature and nurture, and increasingly, to be able to alter our genetic constitution.

Our preoccupation with DNA has already shed light on historical puzzles such as the roots and migrations of ancient peoples, historical demography of cultures and human genetic diversity. We now know that the age-old prejudices of caste, race or royalty, which granted undue importance to biological inheritance, have no substantial genetic basis. The HGP gives the scientific basis for promoting the concept of one race and of equal opportunity for all at the beginning of life. Sadly, this knowledge, on its own, is unlikely to be enough to overcome deeply embedded racial and other prejudices.

Each success of the HGP, from conceptualization through to mapping and DNA sequencing, followed close behind advances in technology and implementation of new strategies, one building on the other. This large-scale, technology-driven approach will epitomize how science is to be conducted in the 21st century. To identify genes involved in disease we will determine the differences in DNA specific to those afflicted. To identify traits unique to *Homo sapiens* we will study differences in DNA between species. To sort through all of this information we will design even more massive computers and newer technologies. And the time between acceptance of a new technology and its application will only become shorter. To keep up with this revolution we may be forced to redesign ourselves using the same tools that brought on the change.

It took less than 50 years from the discovery of the structure of the DNA molecule in 1953 to the cracking of the human genetic code. In the upcoming years, scientists will work vigorously on the HGP to find new genes, to determine the function of the gene products and to apply all of this information to the study of common diseases. While the HGP will not be a cure-all, for some diseases there will be cures in our lifetime. And because of the HGP, preventive medicine through early diagnosis and directive genetic counselling will, over time, move to the forefront of health care, replacing our current reactive system.

Perhaps more difficult than the genetic science will be the social and moral implications of this knowledge. It raises issues of genetic privacy and ownership of our genome, not to mention human cloning, social Darwinism and perhaps even eugenics. In the past, our biological evolution (based on genes) was slow and random caused by the changes in the environment, whereas, our cultural evolution (based on ideas) was fast and purposeful due to frequent formation of new ideas and technologies to disseminate them. In the future, we may be able to catalyze evolution in the same way as cultural change has occurred. It will be the responsibility of all in society to ensure that there is proper education and adequate discussion of these topics to ensure the best decisions are made for our species as a whole as we move forward. Therefore, programs studying ethical, legal and societal implications of genetic information will be as important to success in applying the HGP knowledge as continued support of the science.

Stephen W. Scherer is a Senior Scientist in Genetics and Genomic Biology and Associate Director of The Centre for Applied Genomics at The Hospital for Sick Children in Toronto. He is also Associate Professor in Molecular and Medical Genetics at the University of Toronto. His laboratory has contributed to mapping, sequencing and gene identification studies on human chromosome 7 as part of the HGP. He is currently Chair of the Human Genome Organization (HUGO) Annotation Committee.