Trans. Natl. Acad. Sci. & Tech., Philippines (2001) **23:** 155-161

THE HUMAN GENOME PROJECT: HEALTH PROSPECTS IN THE POST-GENOMIC ERA*

BERNADETTE L. RAMIREZ
Associate Professor, Department of Biochemistry & Molecular Biology, College of Medicine
University of the Philippines, Manila

ABSTRACT

The human genome is the full complement of genetic material in a human cell. In 1990 the United States Department of Energy and the National Institutes of health developed a joint research plan for their genome programs. The goals of the Human Genome Project are the following: 1) genetic and physical mapping of the genome, 2) DNA sequencing, identifying and locating genes, and 3) pursuing further developments in technology and informatics. In addition, the plan emphasizes the continuing importance of the ethical, legal, and social implications of genome research, and it underscores the critical roles of scientific training, technology transfer, and public access to research data and materials.

The public and privately-funded human genome project consortia have today jointly announced completion of a first draft of the human genome sequence by June 2000. Analysis of the sequence so far predicts 38,000 genes, far fewer than the 60-100,000 that the genome had been thought to contain, though the total may rise to some extent as analysis continues. The "post-genomic era", involving the massive challenges of elucidating gene function and uncovering the genetic basis of human variation, has officially begun.

The atlas of the human genome will revolutionize medical practice and biological research into the 21st century and beyond. All human genes will eventually be found, and accurate diagnostics will be developed for most inherited diseases. In addition, animal models for human disease research will be more easily developed, facilitating the understanding of gene function in health and disease.

As research progresses, investigators will also uncover the mechanisms for diseases caused by several genes or by a gene interacting with environmental factors. Genetic susceptibilities have been implicated in many major disabling and fatal diseases including heart disease, stroke, diabetes, and several kinds of cancer. The identification

*Excerpts taken from "To Know Ourselves," a publication of the US Department of Energy and the Human Genome Project.

of these genes and their proteins will pave the way to more-effective therapies and preventive measures. Investigators determining the underlying biology of genome organization and gene regulation will also begin to understand how humans develop from single cells to adults, why this process sometimes goes awry, and what changes take place as people age. New technologies developed for genome research will also find myriad applications in industry, as well as in projects to map (and ultimately improve) the genomes of economically important farm animals and crops.

Genomics has stimulated the growth of several new and exciting down-stream disciplines, namely, functional genomics, proteomics, pharmacogenomics and gene therapy. These new sciences, sitting firmly on the shoulders of genomics and genetics are set to carry on the more practical aspects of its sequencing foundations up to and including a whole new world of drug discovery and therapeutics. It will someday include individualized medicines and discreet gene replacement. The fearsome diseases of cancer, cardiovascular and metabolic disorders, Alzheimer's Diseases, obesity and inherited genetic diseases have been taken on with a bravado borne of dedication, genius and cautious enthusiasm.

INTRODUCTION

The human genome is the full complement of genetic material in a human cell. The genome is distributed among 23 sets of chromosomes. At a more basic level, the genome is DNA, deoxyribonucleic acid, a natural polymer built up of repeating nucleotides, each consisting of a simple sugar, a phosphate group, and one of four nitrogenous bases.

In 1990 the United States Department of Energy and the National Institutes of health developed a joint research plan for their genome programs. Outlined in these programs are specific goals for the next 5 years. Three years later, emboldened by progress that was on track or even ahead of schedule, the two agencies put forth an updated five-year plan. Improvements in technology, together with the experience of three years, allowed an even more ambitious prospect. In broad terms, the revised plan includes goals for 1) genetic and physical mapping of the genome, 2) DNA sequencing, identifying and locating genes, and 3) pursuing further developments in technology and informatics. In addition, the plan emphasizes the continuing importance of the ethical, legal, and social implications of genome research, and it underscores the critical roles of scientific training, technology transfer, and public access to research data and materials. Most of the goals focus on the human genome, but the importance of continuing research on widely studied "model organisms" such as the mouse genome is also recognized.

Some of the central goals for 1993-98 are as follows:

- Complete a genetic linkage map at a resolution of two to five centimorgans by 1995. (This goal was achieved by the September of 1994).
- Complete a physical map at a resolution of 100 kilobases by 1998. By the end of 1995, molecular biologists were halfway to this goal: A physical map was announced with 15,000 sequence-tagged signposts.

- Develop the capacity to sequence 50 million base pairs per year in long continuous segments by 1998.
- Develop efficient methods for identifying and locating known genes on physical maps or sequenced to home in on and ultimately to understand the most important human genes, namely, the ones responsible for serious diseases and those crucial for healthy development and normal functions.
- Pursue technological developments in areas such as automation and robotics.
- Continue the development of database tools and software for managing and interpreting genome data – This is the area of informatics.
- Continue to explore the ethical, legal, and social implications of genome research.

Mapping the Genome

One of the central goals of the Human Genome Project is to produce a detailed "map" of the human genome. There are different kinds of human inheritance patterns. It indicates for each chromosome the location of genes or other "heritable markers," with distances measured in centimorgans, a measure of recombination frequency. By the end of 1994, a comprehensive map was available that included more than 5800 such markers, including genes implicated in cystic fibrosis, myotonic dystrophy, Huntington disease, Tay-Sachs disease, several cancers, and many other maladies. The average gap between markers was about 0.7 centimorgan.

Other maps are known was physical maps, so called because the distances between features are measured not in genetic terms, but in "real" physical units, typically, numbers of base pairs. Methods to produce physical maps of higher resolution are now available.

Sequencing the genome

Ultimately, though, these physical maps and the clones they point to are mere stepping stones to the most visible goal of the genome project, the sequence of three billion characters – A's, T's, C's, and G's – that defines the human species. Included, of course, would be the sequence for every gene, as well as the sequences for stretches of DNA whose functions we don't know yet. As part of the Human Genome Project, millions of base pairs have been sequenced and archived in databases.

The challenge of sequencing the genome is largely one of being able to do the task at a cheaper and faster rate. At the beginning of the Human Genome Project, the cost of sequencing a single base pair was between $2 and $10, and one researcher could produce between 20,000 and 50,000 base pairs of continuous, accurate sequence in a year. Sequencing the genome by the year 2005 would

therefore likely cost $10-20 billion and require a dedicated cadre of at least 5000 workers. Clearly, a major effort in technology development was called for – an effort that would drive the cost well below $1 per base pair and that would allow automation of the sequencing process. From the beginning, therefore, research to develop new technologies, including new cloning vectors, and to establish suitable resources for sequencing, including clone libraries of expressed sequences had received top priority.

Efforts to develop new cloning vectors have been especially productive. YACs remain a classic tool for cloning large fragments of human DNA, but they are not perfect. Some regions of the genome, for example, resist cloning in YACs, and others are prone to rearrangement. New vectors such as bacterial artificial chromosomes (BACs), P1 phages, and P1-derived artificial cloning systems (PACs) have thus been devised to address these problems. These new approaches are critical for ensuring that the entire genome can be faithfully represented in clone libraries, without the danger of deletions, rearrangements, or spurious insertions.

Significant progress is also evident in the development of sequencing technologies, though all of those in widespread current use are still based on methods developed in 1977 by Allan Maxam and Walter Gilbert and by Frederick Sanger and his coworkers. Both methods rely on gel-based electrophoresis systems to separate DNA fragments, and recent advances in commercial systems include increasing the number of gel lanes, decreasing run times, and enhancing the accuracy of base identification. As a result of such improvements, a standard sequencing machine can now turn out raw, unverified sequences of 50,000 to 75,000 bases per day.

Tools of the Trade

A number of innovations have been developed in mapping and sequencing technologies. But several of the central tools of molecular genetics had also been very helpful. One such tool is the class of DNA-cutting proteins known as restriction enzymes. These enzymes, the first of which were discovered in the late 1960s, cleave double-stranded DNA molecules at specific recognition sites, usually four or six nucleotides long.

A second essential tool of modern molecular genetics is gel electrophoresis. It is through this method that DNA fragments of different sizes are most often separated. In classical gel electrophoresis, electrically charged macromolecules are caused to migrate through a polymeric gel under the influence of an imposed static electric field. In time the molecules sort themselves by size, since the smaller ones move more rapidly through the gel than do larger ones. In 1984 a further advance was made with the invention of pulsed-field gel electrophoresis, in which the strength and direction of the applied field is varied rapidly, thus along DNA strands of more than 50,000 base pairs to be separated.

A third necessary tool is DNA "amplification." The classic example is the cloning vector, which may be circular DNA molecules derived from bacteria or

from bacteriophages (viruslike parasites of bacteria), or artificial chromosomes constructed from yeast or bacterial genomic DNA. Another way of amplifying DNA is the polymerase chain reaction, or PCR. This enzymatic replication technique requires that initiators, or PCR primers, be attached as short complementary strands at the ends of the separated DNA fragments to be replicated. An enzyme then completes the synthesis of the complementary strands, thus doubling the amount of DNA originally present. Again and again, the strands can be separated and the polymerase reaction repeated – so effectively, in fact, that DNA can be amplified by 100,000-fold in less than three hours. As with cloning vectors, the result is a large collection of copies of the original DNA fragment.

The Mouse Genome Projeet

The human genome is not so very different from that of the mouse. Obviously, the differences are critical, but so are the similarities. In particular, genetic experiments on other organisms such as the mouse can illuminate much that we could not otherwise learn about homologous human genes that is, genes that are basically the same in the two species.

In some cases, the connection between a newly identified human gene and a known health disorder can be quickly established. More often, however, clear links between cloned genes and human hereditary diseases or disease susceptibilities are extremely elusive. Diseases that are modified by other genetic predispositions, for example, or by environment, diet, and lifestyle can be exceedingly difficult to trace in human families. The same holds for very rare diseases and for genetic factors contributing to birth defects and other developmental disorders. By contrast, disorders such as these can sometimes be followed relatively easily in animal systems, where uniform genetic backgrounds and controlled breeding schemes can be used to avoid the variability that often confounds human population studies. As a consequence, researchers looking for clues to the causes of many complex health problems are focusing more and more attention on model animals systems.

Bioinformatics

Among the less visible challenges of the Human Genome Project is the daunting task of coping with all the sequence data. Appropriate information systems are needed not only during data acquisition, but also for sophisticated data analysis and for the management and public distribution of unprecedented quantities of biological information. The roles of laboratory data acquisition and management systems include the construction of genetic and physical maps, DNA sequencing, and gene expression analysis. These systems typically comprise databases for tracking biological materials and experimental procedures, software for controlling robots or other automated systems, and software for acquiring laboratory data and presenting it in useful form. Among such systems are physical

mapping databases developed at Livermore and Los Alamos, robot control software developed at Berkeley and Livermore, and DNA sequence assembly software developed at the University of Arizona. These systems are the keys to efficient, cost-effective data production in both DOE laboratories and the many other laboratories that use them.

The interpretation of map and sequence data is the job of data analysis systems. These systems typically include task-specific computational engines, together with graphics and user-friendly interfaces that invite their use by biologists and other non-computer scientists. The genome informatics program is the world leader in developing automated systems for identifying genes in DNA sequence data from humans and other organisms, supporting efforts at Oak Ridge National Laboratory and elsewhere. The Oak Ridge-developed GRAIL system is a world-standard gene identification tool. In 1995 alone, more than 180 million base pairs of DNA were analyzed with GRAIL.

The "working draft" of Human Genome Sequence

The public and privately-funded human genome project consortia have today jointly announced completion of a first draft of the human genome sequence by June 2000. This "working draft" is actually far more complete than was envisaged even a short time ago: approximately 24% of the sequence is in finished form (that is, with no gaps and an accuracy of 99.9%) and 50% is close to that state. Astonishingly, 60% of the sequence has been produced within the last six months prior to the announcement. Analysis of the sequence so far predicts 38,000 genes, far fewer than the 60-100,000 that the genome had been thought to contain, though the total may rise to some extent as analysis continues. A related harvest from the genome sequencing effort has been a rich resource of single-nucleotide polymorphisms: sites in the genome where the sequences varies among different individuals by only a single base pair. It is hoped that it may be possible to correlate these variants with differences in human characteristics such as disease susceptibility and drug responses. The SNP project had a target of 100,000 SNPs by 2003, but three times this number have already been discovered and the total is likely to rise to 1 million by the end of this year. The "post-genomic era", involving the massive challenges of elucidating gene function and uncovering the genetics basis of human variation, has officially begun.

Health Prospects in the Post-Genomic Era

The atlas of the human genome will revolutionize medical practice and biological research into the 21st century and beyond. All human genes will eventually be found, and accurate diagnostics will be developed for most inherited diseases. In addition, animal models for human disease research will be more easily developed, facilitating the understanding of gene function in health and disease.

Researchers have already identified single genes associated with a number of diseases, such as cystic fibrosis, Duchenne muscular dystrophy, myotonic dystrophy, neurofibromatosis, and retinoblastoma. As research progresses, investigators will also uncover the mechanisms for diseases caused by several genes or by a gene interacting with environmental factors. Genetic susceptibilities have been implicated in many major disabling and fatal diseases including heart disease, stroke, diabetes, and several kinds of cancer. The identification of these genes and their proteins will pave the way to more- effective therapies and preventive measures. Investigators determining the underlying biology of genome organization and gene regulation will also begin to understand how humans develop from single cells to adults, why this process sometimes goes away, and what changes take place as people age. New technologies developed for genome research will also find myriad applications in industry, as well as in projects to map (and ultimately improve) the genomes of economically important farm animals and crops.

Genomics has stimulated the growth of several new and exciting down-stream disciplines, namely, functional genomics, proteomics, pharmacogenomics and gene therapy. These new sciences, sitting firmly on the shoulders of genomics and genetics are set to carry on the more practical aspects of its sequencing foundations up to and including a whole new world of drug discovery and therapeutics. It will someday include individualized medicines as discreet gene replacement. The fearsome diseases of cancer, cardiovascular and metabolic disorders, Alzheimer's Diseases, obesity and inherited genetic diseases have been taken on with a bravado borne of dedication, genius and cautious enthusiasm. The last few years have seen all of these disorders taken to the clinic in enthusiastic and ambitious genomics-based trials where the assault against them will continue.

Genomics companies are still hard at work identifying sequences and the genes within – an obviously finite endeavor that will eventually be completed. In the meantime, everyone is excited about the upcoming "post-genomic" era and what it will mean to science, business and medicine. The goal, of course, is to connect diseases with the genes responsible. On the surface this may seem like a simple task but it is not.

Diseases symptoms may result from a monogenic defect or miscode within a single gene, e.g., sickle cell anemia – which results from a single miscode for an amino-acid in hemoglobin. Under these conditions, one can hypothesize that replacing the gene sequence that codes for that amino acid will solve the problem. In today's medicine, however, this is still an overwhelmingly difficult task; however, advances in gene therapy promise that such replacements will one day be attainable in a clinically relevant manner.