The Human Genome Project

- and Beyond

Press Conference
April 14, 2003

National Human Genome Research Institute
National Institutes of Health
Department of Health and Human Services
and
Office of Science
U.S. Department of Energy

EMBARGOED
For release after 1 p.m. ET, Monday, April 14, 2003

International Consortium Completes Human Genome Project

All Goals Achieved; New Vision for Genome Research Unveiled

BETHESDA, Md., April 14, 2003 – The International Human Genome Sequencing Consortium, led in the United States by the National Human Genome Research Institute (NHGRI) and the Department of Energy (DOE), today announced the successful completion of the Human Genome Project more than two years ahead of schedule.

Also today, NHGRI unveiled its bold new vision for the future of genome research, officially ushering in the era of the genome. The vision will be published in the April 24 issue of the journal *Nature,* coinciding with the 50th anniversary of *Nature's* publication of the landmark paper by Nobel laureates James Watson and Francis Crick that described DNA's double helix. Dr. Watson also was the first leader of the Human Genome Project.

The international effort to sequence the 3 billion DNA letters in the human genome is considered by many to be one of the most ambitious scientific undertakings of all time, even compared to splitting the atom or going to the moon.

"The Human Genome Project has been an amazing adventure into ourselves, to understand our own DNA instruction book, the shared inheritance of all humankind," said NHGRI Director Francis S. Collins, M.D., Ph.D., leader of the Human Genome Project since 1993. "All of the project's goals have been completed successfully – well in advance of the original deadline and for a cost substantially less than the original estimates."

Aristides Patrinos, Ph.D., director of DOE's Office of Biological and Environmental Research in the Office of Science, said, "Sequencing the human genome was a pioneering venture with risks and uncertainties. But its success has created a revolution – transforming biological science far beyond what we could imagine. We have opened the door into a vast and

complex new biological landscape. Exploring it will require even more creative thinking and new generations of technologies."

The flagship effort of the Human Genome Project has been producing the reference sequence of the human genome. The international consortium announced the first draft of the human sequence in June 2000. Since then, researchers have worked tirelessly to convert the "draft" sequence into a "finished" sequence. Finished sequence is a technical term meaning that the sequence is highly accurate (with fewer than one error per 10,000 letters) and highly contiguous (with the only remaining gaps corresponding to regions whose sequence cannot be reliably resolved with current technology). That standard was first achieved for a human chromosome when a team of British, Japanese and U.S. researchers produced a finished sequence for human chromosome 22 in 1999.

The finished sequence produced by the Human Genome Project covers about 99 percent of the human genome's gene-containing regions, and it has been sequenced to an accuracy of 99.99 percent. In addition, to help researchers better understand the meaning of the human genetic instruction book, the project took on a wide range of other goals, from sequencing the genomes of model organisms to developing new technologies to study whole genomes. As of April 14, 2003, all of the Human Genome Project's ambitious goals have been met or surpassed. **(Table 1-HGP Goals)**

When the Human Genome Project was launched in 1990, many in the scientific community were deeply skeptical about whether the project's audacious goals could be achieved, particularly given its hard-charging timeline and relatively tight spending levels. At the outset, the U.S. Congress was told the project would cost about $3 billion in FY 1991 dollars and would be completed by the end of 2005. In actuality, the Human Genome Project was finished two and a half years ahead of time and, at $2.7 billion in FY 1991 dollars, significantly under original spending projections. **(Table 2-HGP Budget)**

"Never would I have dreamed in 1953 that my scientific life would encompass the path from DNA's double helix to the 3 billion steps of the human genome. But when the opportunity arose to sequence the human genome, I knew it was something that could be done – and that must be done," said Nobel Laureate James D. Watson, Ph.D., president of Cold Spring Harbor Laboratory in Cold Spring Harbor, N.Y. "The completion of the Human Genome Project is a truly momentous occasion for every human being around the globe."

Besides delivering on the stated goals, the international network of researchers has produced an amazing array of advances that most scientists had not expected until much later. These "bonus" accomplishments include: an advanced draft of the mouse genome sequence, published in December 2002; an initial draft of the rat genome sequence, produced in November 2002; the identification of more than 3 million human genetic variations, called single nucleotide polymorphisms (SNPs); and the generation of full-length complementary DNAs (cDNAs) for more than 70 percent of known human and mouse genes.

31 Center Drive MSC 2152 Building 31, Room4B09 Bethesda, MD 20892-2152 TEL 301 402-0911 FAX 301 402-2218

The International Human Genome Sequencing Consortium included hundreds of scientists at 20 sequencing centers in China, France, Germany, Great Britain, Japan and the United States. The five institutions that generated the most sequence were: Baylor College of Medicine, Houston; Washington University School of Medicine, St. Louis; Whitehead Institute/MIT Center for Genome Research, Cambridge, Mass.; DOE's Joint Genome Institute, Walnut Creek, Calif.; and The Wellcome Trust Sanger Institute near Cambridge, England. **(See List-International Human Genome Sequencing Consortium)**

"The enormity of the Human Genome Project is unprecedented in biology. The international vision and collaboration of the scientists involved played a crucial role in the project's success," said Mark Walport, M.D., director designate of The Wellcome Trust, which led the Human Genome Project in the United Kingdom. "The genome is the common thread that connects us all, so it is only fitting that the sequence has been given to us by scientists from all corners of the earth."

All of the sequence data generated by the Human Genome Project has been swiftly deposited into public databases and made freely available to scientists around the world, with no restrictions on its use or redistribution. The information is scanned daily by researchers in academia and industry, as well as by commercial database companies providing information services to biotechnologists.

"From the beginning, one of the operating principles of the Human Genome Project has been that the data and resources it has generated should rapidly be made available to the entire scientific community," said Robert Waterston, M.D., Ph.D., of the University of Washington, Seattle. "Not only does the rapid release of data promote the best interests of science, it also maximizes the benefits that the public receives from such research."

In 1996, at a meeting in Bermuda, Dr. Waterston and John Sulston, Ph.D., then director of the Sanger Centre (now The Wellcome Trust Sanger Institute), led the International Human Genome Sequencing Consortium to adopt the so-called "Bermuda Principles," which expressly call for automatic, rapid release of sequence assemblies of 2,000 bases or greater to the public domain.

Scientists have been quick to mine this new trove of genomic data, as well as to utilize the genomic tools and technologies developed by the Human Genome Project. For example, when the Human Genome Project began in 1990, scientists had discovered fewer than 100 human disease genes. Today, more than 1,400 disease genes have been identified. **(Table 3-Pace of Gene Discovery)**

For scientists seeking to understand the role of genetics in human health and disease, the Human Genome Project's finished sequence represents a significant advance over the "working draft" that was announced in June 2000. The working draft covered 90 percent of the gene-containing part of the sequence, 28% of which had reached finished form, and contained about 150,000 gaps. The finished version of the human genome now contains 99 percent of the gene-containing sequence, with the missing parts essentially contained in less than 400 defined gaps.

These remaining gaps represent regions of DNA in the genome with unusual structures that cannot be reliably sequenced with current technology. These regions, however, appear to contain very few genes. Closing these gaps will require individual research projects and new technologies, rather than industrial-scale efforts of the Human Genome Project. The high-throughput sequencing of the human genome has thus reached its natural conclusion.

"This is the day that our planning group dreamed of," said Bruce Alberts, Ph.D., chairman of the 1988 National Research Council Committee on Mapping and Sequencing the Human Genome, which produced the original recommendations for the Human Genome Project. "And the quality of the sequence would have amazed us. In 1988, we weren't sure that accuracy rates of 99.9 percent were possible, and we were uncertain that continuity over distances of millions of base pairs could be achieved. The finished human sequence is a fabulous outcome. Biomedical researchers now have tremendous foundation on which to build the science and medicine of the 21st century." Dr. Alberts is now the president of the National Academy of Sciences.

In addition to the improved accuracy, the average DNA letter now sits on a stretch of 27,332,000 base pairs of uninterrupted, high-quality sequence – about 334 times longer than the 81,900 base-pair stretch that was available in the working draft. Access to uninterrupted stretches of sequenced DNA can make a major difference to researchers hunting for genes, dramatically cutting the effort and expense required to search regions of the human genome that may contain small and often rare mutations involved in disease.

"The Human Genome Project represents one of the remarkable achievements in the history of science. Its culmination this month signals the beginning of a new era in biomedical research," said Eric Lander, Ph.D., director of the Whitehead-MIT Center for Genome Research. "Biology is being transformed into an information science, able to take comprehensive global views of biological systems. With knowledge of all the components of the cells, we will be able to tackle biological problems at their most fundamental level."

The essentially complete version of the human genome sequence also represents a major boon to the growing field of comparative genomics: researchers are attempting to learn more about human genetic makeup and function by comparing our genomic sequence to that of other organisms, such as the mouse, the rat or even the fruit fly. **(See Comparative Genomics backgrounder.)**

"One of the most powerful tools for understanding our own genome is to study it within the context of a much larger framework. That framework is being created by ongoing efforts to sequence and analyze the genomes of many other organisms," said Richard Gibbs, Ph.D., director of Baylor College of Medicine's Human Genome Sequencing Center. "As we identify the similarities – and the differences – among the genes of mammals and other organisms, we will begin to gain valuable new insights into human evolution, as well as human health and disease."

The impact of the Human Genome Project, however, extends far beyond laboratory analysis. Under the guidance of Dr. Watson, the Human Genome Project became the first large scientific undertaking to dedicate a portion of its budget for research to the ethical, legal and social implications (ELSI) of its work. NHGRI and DOE each set aside 3 to 5 percent of their genome budgets to study how the exponential increase in knowledge about human genetic make-up may affect individuals, institutions and society. An example of how ELSI research has helped to inform public policy is the fact that more than 40 states in the United States have passed genetic non-discrimination bills, many based on model language that grew out of this research. These efforts will be even more crucial in the coming years as the results of genomic research begin to appear in the clinic.

"Achieving the goals of the Human Genome Project is a historic milestone. But this is no time to rest and relax," said Dr. Collins. "With this foundation of knowledge firmly in place, the medical advances promised from the project can now be significantly accelerated."

To spur such acceleration, NHGRI's "A Vision for the Future of Genomics Research" sets forth a series of "Grand Challenges" intended to energize the scientific community in using the newfound understanding of the genome to uncover the causes of disease and to develop bold new approaches to the prevention and treatment of disease. The plan was the outcome of more than a year of intense discussions with nearly 600 scientific and public leaders from government, academia, non-profit organizations and the private sector.

In the publication in *Nature*, the challenges facing genomic research are depicted as a three-story house rising from the foundation of the Human Genome Project. The three floors, representing the three major thrusts of this new vision – Genomics to Biology, Genomics to Health and Genomics to Society – are interconnected by vertical supports, representing computational biology, ELSI, education, training, technology development and resources. **(See attached *Nature* paper)**

Many of the challenges in the vision are aimed at utilizing genome research to combat disease and improve human health. The recommendations include calls for researchers to work toward:

- New tools to allow discovery in the near future of the hereditary contributions to common diseases, such as diabetes, heart disease and mental illness.
- New methods for the early detection of disease.
- New technologies that can sequence the entire genome of any person for less than $1,000.
- Wider access to tools and technologies of "chemical genomics" to improve the understanding of biological pathways and accelerate drug discovery.

31 Center Drive MSC 2152 Building 31, Room4B09 Bethesda, MD 20892-2152 TEL 301 402-0911 FAX 301 402-2218

NHGRI and its partners in genome research have already begun tackling a number of these challenges. For example, in November 2002, a team of researchers from six nations launched the International HapMap Project, an effort to produce a map of common human genetic variations aimed at speeding the search for genes that contribute to cancer, diabetes, heart disease, schizophrenia and many other common conditions. **(See What's Next? backgrounder)**

"The completion of the Human Genome Project should not be viewed as an end in itself. Rather, it marks the start of an exciting new era – the era of the genome in medicine and health," said Dr. Collins. "We firmly believe the best is yet to come, and we urge all scientists and people around the globe to join us in turning this vision into reality."

NHGRI's U.S. partner in the Human Genome Project, DOE, has also developed its own forward-looking plan for genome research. The DOE plan, published in the April 11 issue of the journal *Science,* is focused on understanding the ways in which microbes can provide new opportunities for developing clean energy, reducing climate change and cleaning the environment. To achieve that vision, DOE has begun the "Genomes to Life" program, which will combine research in biology, engineering and computation with the development of novel facilities for high-throughput biology projects.

NHGRI is one of the 27 institutes and centers at the National Institutes of Health, an agency of the Department of Health and Human Services (DHHS). Additional information about NHGRI can be found at its Web site, www.genome.gov.

###

31 Center Drive MSC 2152 Building 31, Room4B09 Bethesda, MD 20892-2152 TEL 301 402-0911 FAX 301 402-2218

National Human Genome Research Institute
National Institutes of Health
Department of Health and Human Services
and
Office of Science
U.S. Department of Energy

International Human Genome Sequencing Consortium

1. Whitehead Institute/MIT Center for Genome Research, Cambridge, Mass., U.S.
2. The Wellcome Trust Sanger Institute, The Wellcome Trust Genome Campus, Hinxton, Cambridgeshire, U.K.
3. Washington University School of Medicine Genome Sequencing Center, St. Louis, Mo., U.S.
4. U. S. Department of Energy Joint Genome Institute, Walnut Creek, Calif., U.S..
5. Baylor College of Medicine Human Genome Sequencing Center, Department of Molecular and Human Genetics, Houston, Tex., U.S.
6. RIKEN Genomic Sciences Center, Yokohama, Japan
7. Genoscope and CNRS UMR-8030, Evry, France
8. GTC Sequencing Center, Genome Therapeutics Corporation, Waltham, Mass., U.S.
9. Department of Genome Analysis, Institute of Molecular Biotechnology, Jena, Germany
10. Beijing Genomics Institute/Human Genome Center, Institute of Genetics, Chinese Academy of Sciences, Beijing, China
11. Multimegabase Sequencing Center, The Institute for Systems Biology, Seattle, Wash., U.S.
12. Stanford Genome Technology Center, Stanford, Calif., U.S.
13. Stanford Human Genome Center and Department of Genetics, Stanford University School of Medicine, Stanford, Calif., U.S.
14. University of Washington Genome Center, Seattle, Wash., U.S.

31 Center Drive MSC 2152 Building 31, Room 4B09 Bethesda, MD 20892-2152 TEL 301 402-0911 FAX 301 402-2218

15. Department of Molecular Biology, Keio University School of Medicine, Tokyo, Japan

16. University of Texas Southwestern Medical Center at Dallas, Dallas, Texas, U.S.*

17. University of Oklahoma's Advanced Center for Genome Technology, Dept. of Chemistry and Biochemistry, University of Oklahoma, Norman, Okla., U.S.

18. Max Planck Institute for Molecular Genetics, Berlin, Germany

19. Cold Spring Harbor Laboratory, Lita Annenberg Hazen Genome Center, Cold Spring Harbor, N.Y., U.S.

20. GBF - German Research Centre for Biotechnology, Braunschweig, Germany

*Sequencing center is no longer in operation.

###

31 Center Drive MSC 2152 Building 31, Room4B09 Bethesda, MD 20892-2152 TEL 301 402-0911 FAX 301 402-2218

EMBARGOED
For release after 1 p.m. ET, Monday, April 14, 2003

Human Genome Project Goals

Area	Goal	Achieved	Date
Genetic Map	2- to 5-cM resolution map (600 – 1,500 markers)	1-cM resolution map (3,000 markers)	September 1994
Physical Map	30,000 STSs	52,000 STSs	October 1998
DNA Sequence	95% of gene-containing part of human sequence finished to 99.99% accuracy	99% of gene-containing part of human sequence finished to 99.99% accuracy	April 2003
Capacity and Cost of Finished Sequence	Sequence 500 Mb/year at < $0.25 per finished base	Sequence >1,400 Mb/year at <$0.09 per finished base	November 2002
Human Sequence Variation	100,000 mapped human SNPs	3.7 million mapped human SNPs	February 2003
Gene Identification	Full-length human cDNAs	15,000 full-length human cDNAs	March 2003
Model Organisms	Complete genome sequences of *E. coli*, *S .cerevisiae*, *C. elegans*, *D. melanogaster*	Finished genome sequences of *E. coli*, *S. cerevisiae*, *C. elegans* *D. melanogaster*, plus whole-genome drafts	April 2003

		of several others, including *C. briggsae*, *D. pseudoobscura*, mouse and rat	
Functional Analysis	Develop genomic-scale technologies	High-throughput oligonucleotide synthesis	1994
		DNA microarrays	1996
		Eukaryotic, whole-genome knockouts (yeast)	1999
		Scale-up of two-hybrid system for protein-protein interaction	2002

Key Definitions

cDNA: cDNA stands for **complementary DNA,** a synthetic type of DNA generated from messenger RNA, or mRNA, the molecule in the cell that takes information from protein-coding DNA - the genes - to the protein-making machinery and instructs it to make a specific protein. By using mRNA as a template, scientists use enzymatic reactions to convert its information back into cDNA and then clone it, creating a collection of cDNAs, or a cDNA library. These libraries are important to scientists because they consist of clones of all protein-encoding DNA, or all of the genes, in the human genome.

cM: cM stands for **centiMorgan,** a unit of genetic distance. Generally, one centiMorgan equals about 1 million base pairs.

Eukaryotic: A eukaryote is a single-celled or multicellular organism whose cells contain a distinct membrane-bound nucleus. If something is described as "eukaryotic," it means that it has cells with membrane-bound nuclei.

Mb: Mb stands for **megabase,** a unit of length equal to 1 million base pairs and roughly equal to 1 cM.

Microarray: Microarrays are devices used in many types of large-scale genetic analysis. They can be used to study how large numbers of genes are expressed as messenger RNA in a particular tissue, and how a cell's regulatory networks control vast batteries of genes simultaneously. In microarray studies, a robot is used to precisely apply tiny droplets containing functional DNA to glass slides. Researchers then attach fluorescent labels to complementary DNA (cDNA) from the tissue they are studying. The labeled cDNA binds to its matched DNA sequence at a specific location on the slide. The slides are put into a scanning

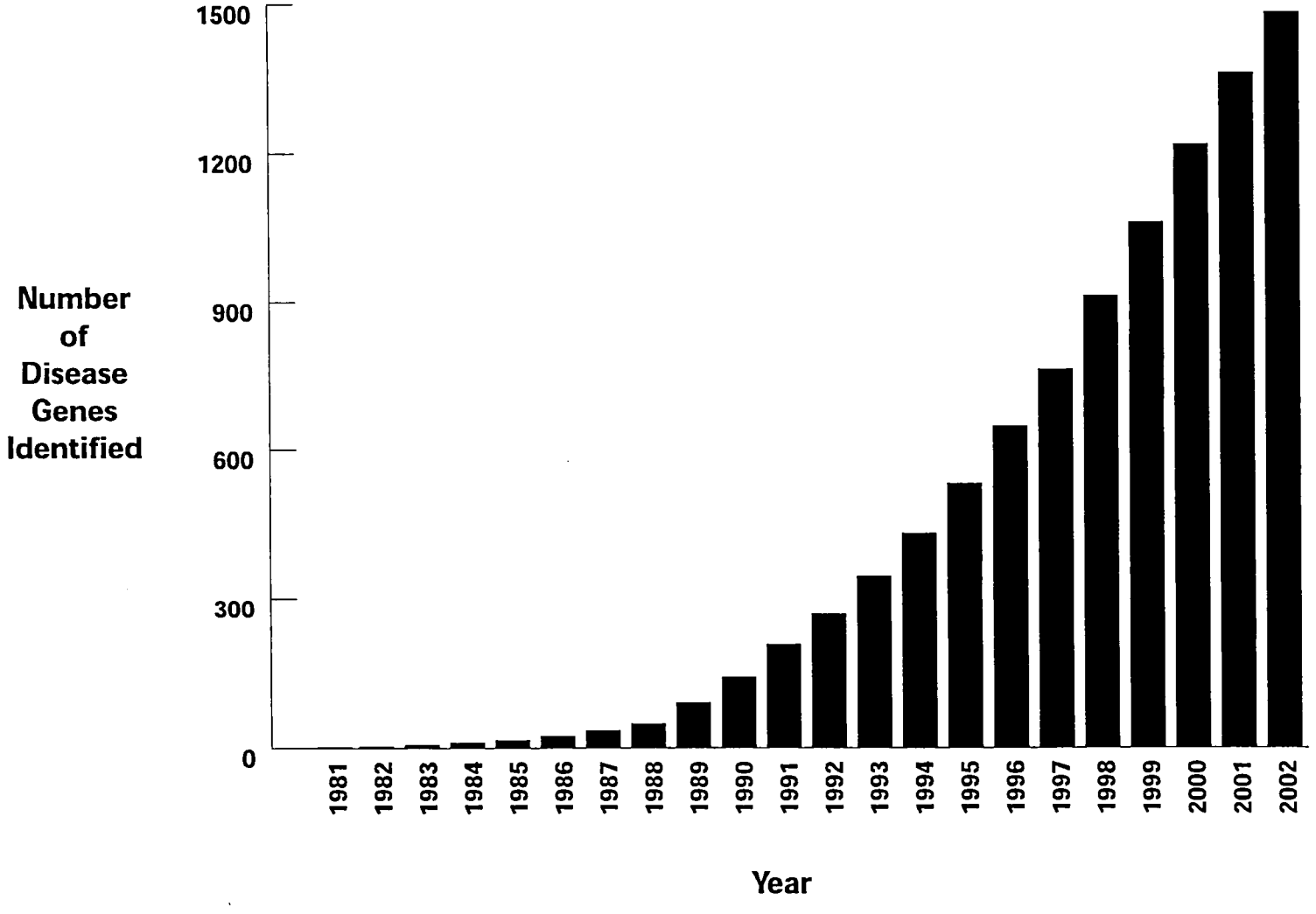
microscope that can measure the brightness of each fluorescent dot. The brightness reveals how much of a specific cDNA fragment is present, an indicator of how active a gene is.

Scientists use microarrays in many different ways. For example, microarrays can be used look at which genes in cells are actively making products under a specific set of conditions, as well as to detect and/or examine differences in gene activity between healthy and diseased cells.

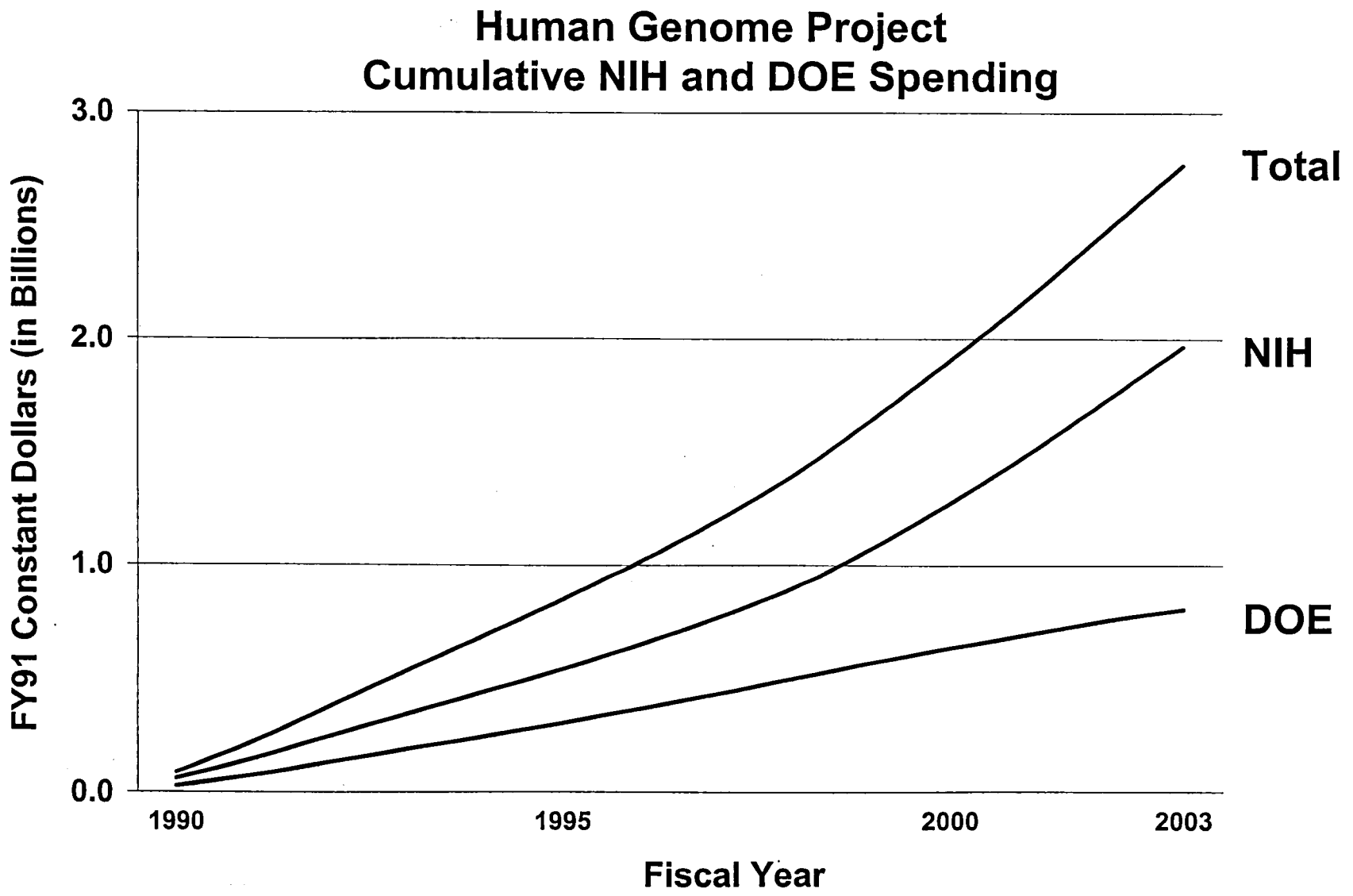
Oligonucleotide: A short polymer of 10 to 70 nucleotides. A nucleotide is one of the structural components, or building blocks, of DNA and RNA. A nucleotide consists of a base chemical – either adenine (A), thymine (T), guanine (G) or cytosine (C) – plus a sugar-phosphate backbone. Oligonucleotides are often used as probes for detecting complementary DNA or RNA because they bind readily to their complements.

SNP: SNP stands for **single nucleotide polymorphism.** SNPs – pronounced "snips" – are common, but minute, variations that occur in the human genome at a frequency of one in every 300 bases. That means 10 million positions out of the 3 billion base-pair human genome have common variations. These variations can be used to track inheritance in families and susceptibility to disease, so scientists are working hard to develop a catalogue of SNPs as a tool to use in their efforts to uncover the causes of common illness like diabetes or heart disease.

STS: STS stands for **sequence tagged site,** a short DNA segment that occurs only once in a genome and whose exact location and order of bases is known. Because each is unique, STSs are helpful in chromosome placement of mapping and sequencing data from many different laboratories. STSs serve as landmarks on the physical map of a genome



Cumulative Pace of Disease Gene Discovery (1981-2002). The number of disease genes identified so far is 1,485. Data provided by Online Mendelian Inheritance in Man.


Human Genome Project
Cumulative NIH and DOE Spending
FY91 Constant Dollars (in Billions)
3.0
2.0
1.0
0.0
1990
1995
2000
2003
Fiscal Year
Total
NIH
DOE