XVIII.
National Human Genome Research Institute

INTRODUCTION

The National Human Genome Research Institute (NHGRI) was established in 1989 to lead the effort of the National Institutes of Health (NIH) in the Human Genome Project (HGP). NHGRI's Division of Extramural Research funds HGP research in laboratories throughout the country. Research on genetic and physical mapping, DNA sequencing, database development, and technology development for genome research, as well as studies of the ethical, legal, and social implications of genetics research, are supported by the extramural arm of NHGRI. In February 1993, the Institute expanded its role at the NIH by establishing the Division of Intramural Research, a cutting-edge program that (a) translates the tools of the HGP into knowledge about human genetic disease and its diagnosis and treatment and (b) serves as the hub for human genetics research at the NIH. NHGRI manages or comanages two NIH-wide scientific service centers that are supported by various NIH Institutes and Centers. The Center for Inherited Disease Research provides high-throughput genotyping services, advice on study design, sophisticated technologies for data warehousing, and database assistance for research efforts attempting to identify gene variants involved in human disease. The NIH Intramural Sequencing Center provides NIH intramural investigators with access to large-scale DNA sequencing and sequence analysis.

NHGRI, through its extramural and intramural research programs, contributes to identification of genes involved in human disease and to study of the functions of these genes and their products. The HGP provides data, material resources, and technology that will improve the ability of scientists to conduct biological research rapidly, efficiently, and cost-effectively. This infrastructure has already dramatically accelerated the study of human inherited disease. In the laboratories of the Division of Intramural Research, with the tools produced by the HGP, scientists are developing and using the most advanced techniques to study the fundamental mechanisms of inherited and acquired genetic disorders.

HIGHLIGHTS OF RECENT SCIENTIFIC ADVANCES RESULTING FROM INTERNATIONAL ACTIVITIES

Progress on Human Genome Project

The HGP is an international research effort to characterize the human genome and the genomes of selected model organisms, through complete mapping and sequencing of the DNA. The goals of the Project are to develop technologies for genomic analysis; examine the ethical, legal, and social implications of research in human genetics; and train scientists to use the tools and resources developed through the Project to pursue biological studies that will improve human health. Begun in October 1990, the HGP is funded in the United States by NHGRI and the U.S. Department of Energy. After the success of the pilot phase, in March 1999, an international consortium launched the full-scale effort to sequence the estimated 3 billion base pairs that make up the human genetic instruction book. The international consortium consists of scientists at 16 institutions in China, France, Germany, Japan, the United Kingdom, and the United States.

The production of human genome sequence skyrocketed in fiscal year 2000 (FY 00). During this time, the international sequencing consortium produced 1,000 bases a second of raw sequence 7 days a week, 24 hours a day. On June 26, 2000, leaders of the public HGP and Celera Genomics Corporation announced that both had successfully completed the production of a "working draft" of the human genome. This historic scientific milestone was announced with President Bill Clinton at a White House event and included a satellite link to Prime Minister Tony Blair and genome leaders in the United Kingdom who contributed significantly to the public effort.

The working draft is substantially closer to the ultimate finished form than the consortium expected at this stage. Approximately 50% of the genome sequence is nearly finished, and 25% of it is completed. The average accuracy of all of the DNA sequence in this assembly is 99.9%.

The public HGP and Celera Genomics Corporation used different but complementary approaches to sequencing the human genome. The public project used a "hierarchical shotgun" approach in which individual large DNA fragments of known position are subjected to shotgun sequencing. Celera Genomics Corporation used a "whole-genome shotgun" approach, in which the entire genome is shredded into small fragments that are sequenced and put back together on the basis of sequence overlaps.

Public and private research teams will submit their genomic data simultaneously for publication in FY 01. After publication, both sets of teams will join together for a workshop to compare the two sequencing methods. The intense phase of analyzing the sequence for gene content and a host of other biological features is under way. A particularly useful data set, representing the current "best view" of the genome sequence, is the so-called Golden Path, which can be found at http://genome.cse.ucsc.edu/.

DNA Sequencing

The current goal of the HGP is completion of the first high-quality reference human DNA sequence by 2003 or possibly earlier. The working draft completed in FY 00, while not as refined as the final version, provides the community with sequence covering most of the genome and represents the raw data needed to find most of the human genes. The final human genome sequence will be highly refined, that is, 99.99% accurate. The only gaps in the finished sequence will be those that cannot be closed by exist-

ing technology. All gaps will be annotated to indicate the size, the orientation of the adjoining sequence, and the reason why the gap was not filled.

The international sequencing programs continue to submit all sequence assemblies of 1–2 kilobases or more, within 24 hours of their generation, into public databases where they are immediately and freely released to the world, with no restrictions on use or redistribution. The information is scanned daily by scientists in academia and industry, as well as by commercial database companies providing information services to biotechnologists. A list of links to a number of important web sites that contain information about the human genome sequence, other genome sequences, and other relevant genomic information can be found at http://www.nhgri.nih.gov/genome_hub.html.

On March 14, 2000, President Clinton and British Prime Minister Blair endorsed this standard of practice when they announced a statement of principle to ensure that discoveries from the human genome are used to advance human health. Their joint statement applauded researchers who have made their human genome sequence data freely available to the global scientific community. The statement also acknowledged the importance of intellectual property protection as an incentive for the development of important, new gene-based health care products.

Unraveling Genetic Code of Human Chromosomes 21 and 22

In December 1999, an international team of researchers at the Sanger Centre, Cambridge, England, the University of Oklahoma, Norman, Washington University, St. Louis, Missouri, and Keio University, Tokyo, Japan, reported for the first time, the sequencing of the 33.5 million base pairs of chromosome 22. The sequence included the longest continuous stretch of DNA ever assembled, at more than 23 million base pairs.

Just a few months later, in May 2000, scientists in Germany and Japan published the finished genetic sequence of human chromosome 21. Studying the organization of the genes on chromosome 21 and how they and their protein products function should help scientists to find clues about Down syndrome, as well as other disorders that have been linked to this chromosome.

With the complete DNA sequence of chromosomes 21 and 22 now in hand, scientists can begin to study structural similarities between and among chromosomes, as well as shared sequences. Seeing the structure and organization of chromosomes 21 and 22 at the base pair level for the first time immediately suggested new experiments and avenues of research to be pursued. It permits scientists to begin to understand where genes are located on chromosomes, how they express themselves, how deletions that give rise to disease-causing mutations occur, and how chromosomes are duplicated and inherited.

Single Nucleotide Polymorphisms: New Tools for Tracing Inherited Diseases

A key aspect of research in genetics is associating sequence variations with heritable phenotypes. The most common variations are single nucleotide polymorphisms (SNPs), which occur approximately once in every 100–300 bases. Because SNPs are expected to facilitate large-scale association genetics studies, there has recently been great interest in SNP discovery and detection. The identification of SNPs has accelerated dramatically in FY 00, largely because of the availability of the working draft sequence of the human genome.

In FY 99, NHGRI organized the establishment of the DNA Polymorphism Discovery Resource. To maximize the chances of discovering common DNA sequence polymorphisms, 450 DNA samples were collected, under strict ethical guidelines, from anonymous unrelated U.S. residents who have ancestors from one or more major geographic regions of the world—Africa, the Americas, Asia, and Europe. The DNA Polymorphism Discovery Resource is now the major resource being used to look for SNPs. NHGRI has funded studies to allow researchers to look for SNPs in a common set of samples. This permits the accumulation of a haplotype data set that can then be used in association studies to identify inherited disease risks.

This effort is complemented by the SNP Consortium, a nonprofit entity with the mission to develop a high-quality SNP map of the human genome and to make the information related to these SNPs available to the public without intellectual property restrictions. Consortium members include the medical research charity the Wellcome Trust; Motorola, Incorporated; IBM; Amersham Pharmacia Biotech; and 10 pharmaceutical companies, including AstraZeneca PLC, Aventis Pharma, Bayer AG, Bristol-Myers Squibb Company, Hoffmann-La Roche, Glaxo Wellcome PLC, Novartis, Pfizer Incorporated, Searle (now part of Pharmacia), and SmithKline Beecham PLC. Academic centers involved in SNP identification and analysis include Stanford Human Genome Center, California; Whitehead Institute for Biomedical Research, Cambridge, Massachusetts; Washington University School of Medicine, St. Louis, Missouri; Cold Spring Harbor Laboratory, Cold Spring, New York; and the Wellcome Trust's Sanger Centre, Cambridge, England.

In July 2000, the HGP and the SNP Consortium announced an international collaboration to accelerate the construction of a higher-density SNP map and enhance the utility of the human working draft sequence. At the same time, the data generated will help to improve the working draft itself. Three academic genome research centers—Whitehead Institute for Biomedical Research, Washington University School of Medicine, and the Sanger Centre—will participate in this collaboration.

Through this collaboration, the SNP Consortium will be able to contribute about three times as many SNPs to the public domain as otherwise would have been possible under the consortium's original scientific plan. The plan was to identify 300,000 SNPs and map at least 150,000 of these SNPs, evenly distributed throughout the genome. An exponential increase in the amount of human genetic sequence data that has recently become available from the HGP has enabled the consortium to proceed at a much faster pace than was originally envisioned.

SUMMARY OF INTERNATIONAL PROGRAMS AND ACTIVITIES

Extramural Programs

In addition to support for sequencing of the part of the human genome that is being explored by U.S. scientists, NHGRI continues to fund a portion of an international consortium involved in the *Saccharomyces*

Genome Deletion Project. This project aims (1) to generate a complete set of yeast mutants resulting from deletion of genes and (2) to assign functions to the genes by studying the mutants. The strains with deletion(s) are being generated by a consortium of nine European and nine North American laboratories. The set of mutants is nearly complete, and they are available through several distributors, including one in Europe.

International Databases

The NIH provided the sole support for the following international databases: Mouse Genome Database; SacchDB (yeast, *Saccharomyces cerevisiae*); OMIM (Online Mendelian Inheritance in Man); and GeneClinics (a medical genetics knowledge base). Flybase (*Drosophila melanogaster*) is supported by the NIH and the British Medical Research Council.

The Mouse Genome Database and Flybase are maintained locally by institutions in Australia, France, Japan, and the United Kingdom. A mirror Flybase site is maintained locally in Israel.

International Meetings

NHGRI provided support for the following international meetings in FY 00:

- 13th International Mouse Genome Meeting, in Philadelphia, Pennsylvania, in October–November 1999;
- 6th International Strategy Meeting on Human Genome Sequencing, in Walnut Creek, California, in January 2000;
- 8th International Workshop on Chromosomes in Solid Tumors, at the University of Arizona, Tucson, in January–February 2000;
- meetings of the five largest international genome-sequencing centers, in Walnut Creek, California, in January, and in Houston, Texas, in April 2000;
- National Hemophilia Foundation, in Washington, D.C., in March 2000;
- How Many SNPs Are Needed for Disease Gene Mapping? in Bethesda, Maryland, in March 2000;
- 1st NIH Conference on Holoprosencephaly, in Washington, D.C., in April 2000;
- 2nd Mouse Follow-up Meeting—Priority Setting for Mouse Genomics and Genetics Resources, in Bethesda, Maryland, in May 2000;
- Genome Mapping, Sequencing, and Biology Meeting, at Cold Spring Harbor Laboratory, Cold Spring, New York, in May 2000;
- 7th International Strategy Meeting on Human Genome Sequencing, at Cold Spring Harbor Laboratory, Cold Spring, New York, in May 2000;
- 3rd Annual Meeting of the American Society of Gene Therapy, in Denver, Colorado, in May–June 2000;
- Trace Repository Workshop, in Bethesda, Maryland, in July 2000;
- 8th International Strategy Meeting on Human Genome Sequencing, in Evry, France, in September 2000; and
- Report of the 1st Community Consultation on Responsible Collection and Use of Samples for Genetic Research, in Bethesda, Maryland, in September 2000.

During FY 00, NHGRI staff met with representatives of government, industry, and academic institutions from various countries, including Canada, Egypt, Finland, France, Greece, Iceland, India, Japan, Switzerland, and the United Kingdom.

Intramural Programs and Activities

West African Origins of Type 2 Diabetes Mellitus in African Americans

During the past several years, the NIH Office of Research on Minority Health has supported innovative joint research by investigators from Howard University, Washington, D.C., and scientists in the intramural research program of NHGRI. The collaboration involves the study of genetic risk factors for type 2 diabetes mellitus in African Americans.

There is a high frequency of environmental risk factors for type 2 diabetes in the African-American population. Thus it is more productive to study genetic risk factors in West Africans, because they are thought by many anthropologists to be the founding population of modern African Americans and to have fewer dietary and nutritional confounding variables. To establish recruitment sites for the study, five sites from a total of 24 applications were selected through a peer-review process: two sites in Ghana and three in Nigeria. Because of logistic challenges involved in doing a study of this type in West Africa, the study was planned in stages, to allow assessment of the ability to recruit appropriate patients at the site; collect blood, urine, and other clinical data; and successfully send the samples and data to the coordinating center at Howard University. The 1-year pilot project met its goal of recruiting 15 affected sibling pairs per site. On the basis of this experience, a full-scale study was implemented in September 1998, with an anticipated total of 400 affected sibling pairs and 200 spouse control subjects from West Africa by the end of the study period.

By the end of FY 00, nearly all of the 400 affected sibling pairs were recruited, along with 200 unaffected spouse control subjects. The DNA from this group was submitted to the Center for Inherited Disease Research, Johns Hopkins University, Baltimore, Maryland, for a genome-wide scan. In addition, as a result of this study, a pilot study was undertaken at the University of Ibadan, Nigeria, to assess community attitudes toward informed consent.

The study has not only started to yield high-quality data, but has assisted in the recruitment of several expert scientists to the center at Howard University.

Hereditary Prostate Cancer Linkage Analysis

Scientists at the University of Tampere and Tampere University Hospital, Finland, are performing a linkage analysis of families with hereditary prostate cancer, in Finland, Iceland, Sweden, and the United States. A report by the group indicates the possibility of linkage of a locus for prostate cancer to a region of chromosome 1 (*Science,* November 1996). A second locus on the X chromosome has also been mapped by linkage analysis (*Nature Genetics,* October 1998).

These results were followed up in FY 00 by intensive linkage analyses of additional families to markers in these regions and in other regions that showed some evidence of linkage in the initial genome scan. Families from several regions of Finland, Sweden, and the United States have been recruited and genotyped. Association analyses had also been performed on data from Finland, Iceland, and the United States. Efforts are also under way to develop additional family resources for this project. These investigators joined the International Consortium for Prostate Cancer to try to localize prostate cancer loci more rapidly, and a meta-analysis report from this consortium describing the evidence for linkage to the chromosome 1 locus in a very large combined data set was pub-

lished in *Human Genetics* (October 2000). This is an ongoing project to identify and clone genes for prostate cancer. A linkage genome-wide scan of families from Finland, Sweden, and the United States, including a significant number of African-American families, started in FY 00 and will continue into FY 01.

Hereditary Breast Cancer

In a joint project led by researchers at NHGRI, researchers at Turku University Hospital, Tampere University, Tampere University Hospital, and Helsinki University Central Hospital, Finland, University Hospital, Reykjavik, Iceland, and University Hospital, Lund, and University Hospital, Umeå, Sweden, have found evidence of a new gene that appears to increase susceptibility to hereditary breast cancer. The study examined women who live in Nordic countries and who have three or more female family members with breast cancer. The finding may help to explain why some women with a family history of hereditary breast cancer are at particularly high risk of developing the potentially fatal disease, even when they lack mutations in two previously identified breast cancer susceptibility genes, BRCA1 and BRCA2. This work resulted in a publication in the *Proceedings of the National Academy of Sciences* (August 2000). Work that will continue into FY 01 includes linkage studies of other families that do not appear to be linked to any of these three regions, in a search for additional breast cancer susceptibility loci.

Finnish-U.S. Investigation of Type 2 Diabetes Mellitus

The Finnish-U.S. Investigation of Type 2 Diabetes Mellitus aims to identify susceptibility genes for type 2 diabetes and for the related intermediate quantitative traits in a Finnish population. A genome scan for type 2 diabetes and related phenotypes at an average marker density less than 10 centimorgans has been completed on almost 3,600 individuals. Those results have pointed to susceptibility genes on chromosomes 6, 11, 20, and 22. The focus of the project has now turned toward optimizing high-throughput genotyping of densely spaced SNPs in the candidate intervals, as well as sampling candidate genes in these intervals to look for functionally significant variants. The project uses denaturing high-performance liquid chromatography for variant detection and mass spectrometry for high-throughput genotyping of SNPs. The ultimate intent is the positional cloning of diabetes susceptibility genes.

Tissue Microarray Technology

NHGRI scientists, in collaboration with pathologists at the University of Switzerland, Basel, developed a new tissue microarray (tissue chip) technology for high-throughput molecular profiling of very large numbers of tissue specimens.

Now that the human genome sequence has been unraveled, researchers will have a rich resource of genes and gene products to analyze for their involvement in all the hundreds of individual cell types in normal tissues, in developing and differentiating tissues, and in cancer and other diseases. Compared with the research in the "pregenomic" era, research will now need to be adapted to the high-throughput genomic scale, and tissue microarray technology enables this adaptation. Up to 1,000 robotically arrayed tissue specimens can be analyzed in a single experiment, to catalog the involvement of genes and proteins in disease development. This technology enables production of thousands of such replicate microarray slides, and each can be studied with different probes and antibodies. This NHGRI-developed technology is now finding widespread use in both basic and clinical cancer research, as well as in the development of new diagnostics and therapeutics. For example, the National Cancer Institute is using this technology to develop microarray slides from cancer tissues and to make these available to cancer researchers across the country.

Oral Clefts

NHGRI collaborates on a study of the genetics of oral clefts (cleft lip, cleft palate, or both) with investigators at Ibn Al-Nafees Hospital, Damascus, Syria. Researchers obtain family history, clinical data, and blood samples from persons whose families have several members affected with oral clefts, without the presence of a well-known genetic syndrome. DNA from the blood samples will be genotyped, and NHGRI scientists will perform statistical analyses by using these data, together with family history and clinical information, to determine whether there is evidence of genetic susceptibility to oral clefts. For any genetic areas tentatively identified, there will be additional investigation with the use of a dense map of genetic markers and further statistical analyses. The chromosomal regions most likely to contain an area that increases risk for oral clefts will be investigated with molecular genetic techniques designed to clone and sequence the gene in question. Several hundred persons have been studied in Syria, and genotyping and linkage analysis of the first set of families has been completed for a genome-wide set of markers. Regions with suggestive evidence of linkage are being studied with fine-mapping techniques in the original set of families and in a new set of recently recruited families. These analyses and the collaborative data collection in Syria will continue into FY 01.

Common Forebrain Defect From Gene Mutation

An international team led by scientists at NHGRI located one of the genes that can cause holoprosencephaly, the most common structural defect of the developing forebrain in humans. It results in varying degrees of mental retardation. The finding suggests that the gene, which produces TG-interacting factor (TGIF), plays an important role in the brain's separation into left and right hemispheres during fetal development. The TGIF gene is the fourth found in humans to be involved in holoprosencephaly.

Genetic Signature for Malignant Melanoma

Researchers at Queensland Institute of Medical Research, Australia, and Hewlett-Packard Laboratories, Haifa, Israel, and in the United States, led by scientists at NHGRI, discovered a genetic "signature" that may help to explain how malignant melanoma can spread to other parts of the body. Using gene expression profiling, the researchers were able to find a genetic signature—a set of differences in genes—that for the first time divided patients with advanced melanoma into subgroups. Such classification of cancer on a molecular level offers the possibility of more accurately determining the prognosis of a particular patient's tumor on the basis of his or her genetic makeup. It also offers the potential of tailoring therapies to the individual.

Linkage Analysis of Colorectal Cancer

In collaboration with the University of Pisa, Italy, researchers have identified a large pedigree that exhibits high frequency of colorectal cancer across five generations. The cancer fits the criteria for hereditary nonpolyposis colorectal cancer (HNPCC) but does not appear to have mutations in several known genes for HNPCC. This family has been genotyped for markers in the regions of all other known genes for HNPCC, and linkage analyses have been performed. These analyses have shown evidence that the colorectal cancer gene in this family is not linked to these other known loci. This family has been typed in a genome-wide screen to look for a new gene predisposing to colorectal cancer. Analyses of these genome-screen data will continue into FY 01.

Molecular Pathogenesis of Chromosome 16 Inversion in Human Leukemia

Chromosome 16 inversion is one of the most common chromosome abnormalities in human acute myeloid leukemia. A fusion gene between the core-binding factor B (CBFB) gene and the myosin heavy-chain 11 (MYH11) gene is generated by this inversion. Using transgenic mouse models, scientists at NHGRI demonstrated recently that CBFB-MYH11 is necessary, but not sufficient, for leukemogenesis and that additional genetic changes are needed for full leukemic transformation. Such cooperating genetic changes in mice are being identified with a retroviral insertional mutagenesis approach. CBFB encodes a transcription factor that associates with other nuclear proteins and regulates expression of target genes. Identification of downstream target genes of CBFB will lead to better understanding of its role in leukemogenesis and may lead to new diagnostic and therapeutic targets. In collaboration with researchers at Westfälische Wilhelms-Universität, Münster, Germany, the project used the cDNA (complementary DNA) microarray technology to analyze gene expression changes in leukemic cells harboring the chromosome 16 inversion and identified potential target genes. In addition, NHGRI researchers, in collaboration with researchers at Western Australia Institute for Medical Research, Perth, and Kyoto University, Japan, are using novel transgenic mouse models to understand the function of CBFB in normal hematopoiesis and the functional domains of CBFB-MYH11 in leukemogenesis.

Genetic Analysis of Attention-Deficit Hyperactivity Disorder

An international collaboration with scientists at the University of Antioquia, Medellín, Colombia, was established in FY 00, to study the genetics of attention-deficit hyperactivity disorder (ADHD). To study the hypothesis that ADHD is a genetically influenced brain disorder, a genome-wide search for loci linked to ADHD is being undertaken. To this end, 100 densely affected multigenerational Hispanic families will be recruited to the study. Participants will have a battery of psychological tests and will have blood drawn for the linkage analysis and positional cloning studies that will be used to search for genes associated with ADHD. Preliminary analysis of large, densely affected pedigree structures from Colombia showed that there is sufficient power to detect genes of moderate-to-large effect, even in the presence of heterogeneity. This study will continue into FY 01.

blank