Genetics Basics: www.ornl.gov/hgmis/project/info.html

Human Genome news

Linking Interdisciplinary Contributors and Users of Genome Data and Resources

U.S. Department of Energy Office of Biological and Environmental Research *www.ornl.gov/hgmis/project/about.html*

Human Genome Project Fact Sheet

May 2001

ISSN: 1050–6101
Excerpts from Vol. 11(1–2)

On the Shoulders of Giants: Private Sector Leverages HGP Successes

Data, Technologies Catalyze a New, High-Profile Life Sciences Industry

The deluge of data and related technologies generated by the Human Genome Project (HGP) and other genomic research presents a broad array of commercial opportunities. Seemingly limitless applications cross boundaries from medicine and food to energy and environmental resources, and predictions are that life sciences may become the largest sector in the U.S. economy.

Established companies are scrambling to retool, and many new ventures are seeking a role in the information revolution with DNA at its core. IBM, Compaq, DuPont, and major pharmaceutical companies are among those interested in the potential for targeting and applying genome data.

In the genomics corner alone, dozens of small companies have sprung up to sell information, technologies, and services to facilitate basic research into genes and their functions. These new entrepreneurs also offer an abundance of genomic services and applications, including additional databases with DNA sequences from humans, animals, plants, and microbes.

Other applications include gene fragments to use for drug development and target identification and evaluation, identification of candidate genes, and RNA expression information revealing gene activity. Products include protein profiles; particular genotypes associated with such specific medically important phenotypes as disease susceptibility and drug responsiveness; hardware, software, and reagents for DNA sequencing and other DNA-based tests; microarrays (DNA chips) containing tens of thousands of known DNA and RNA fragments for research or clinical use; and DNA analysis software.

From the start, HGP planners anticipated and promoted the private sector's participation in developing and commercializing genomic resources and applications. The HGP's successes in establishing an infrastructure and funding high-throughput technology development are giving rise to commercially viable products and services, with the private sector now taking on more of the risk.

A Public Legacy

Substantial public-sector R&D investment often is needed in feasibility demonstrations before such start-up ventures as those by Celera Genomics, Incyte, and Human Genome Sciences can begin. In turn, these companies furnish valuable commercial services that the government cannot provide, and the taxes returned by their successes easily repay fundamental public investments. Following are a few key public R&D contributions that made some current genomics ventures commercially feasible. These examples describe DOE investments, but substantial commitments by NIH and the Wellcome Trust in the United Kingdom were equally important.

Scientific Infrastructure. The scientific foundation for a human genome initiative existed at the national laboratories before DOE established the first genome project in 1986. Besides expertise in a number of areas critical to genomic research, the laboratories had a long history of conducting large multidisciplinary projects.

Genomic Science and Pioneering Technology. GenBank, the world's DNA sequence repository, was developed at Los Alamos National Laboratory (LANL) and later transferred to the National Library of Medicine. Chromosome-sorting capabilities developed at LANL and Lawrence Livermore National Laboratory enabled the development of DNA clone libraries representing the individual

Contents

On the Shoulders of Giants.....1
HGP and Private Sector.....2
Challenges for the Future.....3
Gene Patenting Update.....3
HGP Funding Levels.....3
HGP FAQs.....4
- When is a Genome Completely Finished?.....4
- Whose Genomes?.....4
- Draft vs Finished Sequence.....4
- Why DOE?.....4
- What's the Next Step.....4

Contact Information.....4

Some Applications of Genomic Data

- **Clinical medicine.** Many more individualized diagnostics and prognostics, drugs, and other therapies.
- **Agriculture and livestock.** Hardier, more nutritious, and healthier crops and animals.
- **Industrial processes.** Cleaner, more efficient manufacturing in such sectors as chemicals, pulp and paper, textiles, food, fuels, metals, and minerals.
- **Environmental biotechnology.** Biodegradable products, new energy resources, environmental diagnostics, and less hazardous cleanup of mixed toxic-waste sites.
- **DNA fingerprinting.** Identification of humans and other animals, plants, and microbes; evolutionary and human anthropological studies; and detection of and resistance to harmful agents that might be used in biological warfare.

More information
www.ornl.gov/hgmis/project/benefits.html and *www.ornl.gov/hgmis/elsi/elsi.html*

(see Private Sector, p. 2)

HGP and the Private Sector: Rivals or Partners?

With the June 26 announcement by the publicly funded Human Genome Project (HGP) and Celera Genomics that the draft sequence of the human genome was essentially complete, the complementary aspects of the public and private sectors' sequencing projects were realized.

Since spring 1998, when Celera Genomics announced its sequencing goal, other private companies also have declared their intention to sequence or map genomic regions to varying degrees. Some people questioned whether the HGP and the private sector were duplicating work, and they wondered who would "win" the race to sequence the human genome. Although the HGP and private companies do have overlapping sequencing goals, their "finish lines" are different because their ultimate goals are not the same.

In a sense, through its policy of open data release, the HGP has all along facilitated the research of others. Additionally, the HGP funds projects at small companies to devise needed technologies. DOE, NIH, the National Institute for Standards and Technology, and other governmental funding sources also are supporting further application and commercialization of HGP-generated resources.

HGP products have spurred a boom in such spin-off programs as the NIH Cancer Genome Anatomy Project and the DOE Microbial Genome Program. Genomes of numerous animals, plants, and microbes are being sequenced, and the number of private endeavors is increasing. Technology transfer from developers to users and participation in collaborative, multidisciplinary projects closely unite researchers at academic, industrial, and governmental laboratories.

Scientific vs Commercial Goals

The HGP's commitment from the outset has been to create a scientific standard (an entire reference genome). Most private-sector human genome sequencing projects, however, focus on gathering just enough DNA to meet their customers' needs—probably in the 95% to 99% range for gene-rich, potentially lucrative regions. Such private data continue to be enriched greatly by accurate free public mapping (location) and sequence information.

Celera's shotgun sequencing strategy, for example, creates millions of tiny fragments that must be ordered and oriented computationally using HGP research results. Most data at Celera, Incyte, and other genomics information–based companies are proprietary or available only for a fee. In addition, companies are filing numerous patent applications to stake early claims to genes and other potentially important DNA fragments (see p. 3).

More than the Reference Sequence

DNA sequencing will continue to be a major emphasis for the foreseeable future as gene sequences are surveyed

Private Sector *(from p. 1)*

chromosomes. These libraries were a crucial resource in genome sequencing.

Sequencing Strategies. When the HGP was initiated, vital automation tools and high-throughput sequencing technologies had to be developed or improved. The cost of sequencing a single DNA base was about $10 then; today, sequencing costs have fallen about 100-fold to $.10 to $.20 a base and still are dropping rapidly.

DOE-funded enhancements to sequencing protocols, chemical reagents, and enzymes contributed substantially to increasing efficiencies. The commercial marketing of these reagents has greatly benefitted basic R&D, genome-scale sequencing, and lower-cost commercial diagnostic services..

A Successful Transformation

These successes transferred much of the repetitive labor from humans to automated machines. In addition, new software for data processing both alleviated and sped human decision making. Over the last decade, advances in instrumentation, automation, and computation have transformed the entire process. Further innovations, however, still are needed for completing many large sequences and increasing the effectiveness of sequencing. *[Denise Casey (HGMIS) and Marvin Stodolsky (DOE)]* ♦

Sequencing Technologies, Biological Resources

Other major factors in cost and time reduction are greatly improved sequencing instruments and efficient biological resources such as the following:

- DOE-funded research on capillary-based DNA sequencing contributed to the development of the two major sequencing machines now in use. The core optical system concept of the Perkin-Elmer 3700 sequencing machine (used by Celera and others) was pioneered with DOE support. The instrumentation concepts that matured as the MegaBACE sequencer were pioneered by Richard Mathies (University of California, Berkeley). The DOE JGI chose this sequencing hardware platform after competitive trials.
- DNA sequencing originally was done with radiolabeled DNA fragments. Today, DOE improvements to fluorescent dyes decrease the amount of DNA needed and increase the accuracy of sequencing data.
- Bacterial artificial chromosome (BAC) clones, developed in the DOE program, became the preferred starting resource in sequencing procedures because of their superior stability and large size. A critical component of public- and private-sector sequencing, BACs were used to assemble both the draft and final human DNA reference sequences.
- Further extending the usefulness of BACs, the DOE HGP funded the production of sequence tag connectors (STCs) from BAC ends. This early information enabled the selection of optimal BACs for complete sequencing, thus saving time and money. STC use for the HGP was advocated by Craig Venter and Nobelist Hamilton Smith (both at Celera), and Leroy Hood (now at the Institute for Systems Biology)

across various populations. Both the DOE and NIH genome programs are continuing to support the development of fully integrated and innovative approaches to rapid, low-cost sequencing.

Other near-term HGP goals from the latest 5-year plan are to enhance bioinformatics (computational) resources to support future research and commercial applications. The HGP also aims to explore gene function through comparative mouse-human studies, train future scientists, study human variation, and address critical societal issues arising from the increased availability of human genome data and related analytical technologies. ♦

Links to Draft Data:

- ***www.ornl.gov/hgmis/project/journals/sequencesites.html***

Sites with assembled human genome data (including browsing tools), other research sites, *Nature* and *Science* papers, insights into the data, and press releases

U.S. Human Genome Project Funding ($ Millions)

FY	DOE	NIH	U.S. Total
1987	5.5	0	5.5
1988	10.7	17.2	27.9
1989	18.5	28.2	46.7
1990	27.2	59.5	86.7
1991	47.4	87.4	134.8
1992	59.4	104.8	164.2
1993	63.0	106.1	169.1
1994	63.3	127.0	190.3
1995	68.7	153.8	222.5
1996	73.9	169.3	243.2
1997	77.9	188.9	266.8
1998	85.5	218.3	303.8
1999	89.9	225.7	315.6
2000	88.9	271.7	360.6
2001	86.4	308.4	394.8

Challenges for the Future: What We Still Don't Know

- Gene number, exact locations, and functions
- Gene regulation
- DNA sequence organization
- Chromosomal structure and organization
- Noncoding DNA types, amount, distribution, information content, and functions
- Coordination of gene expression, protein synthesis, and post-translational events
- Interaction of proteins in complex molecular machines
- Predicted vs experimentally determined gene function
- Evolutionary conservation among organisms
- Protein conservation (structure and function)
- Proteomes (total protein content and function) in organisms
- Correlation of SNPs (single-base DNA variations among individuals) with health and disease
- Disease-susceptibility prediction based on gene sequence variation
- Genes involved in complex traits and multigene diseases
- Complex systems biology including microbial consortia useful for environmental restoration
- Developmental genetics, genomics

Gene Patenting Update: U.S. PTO Tightens Requirements

Massive amounts of data flowing from the Human Genome Project and other genomics projects have stimulated an avalanche of applications to the U.S. Patent and Trademark Office (PTO) for patents on genes and gene fragments. Some 3 million ESTs (fragments that identify pieces of genes) and thousands of other partial and whole genes are included within pending patents. This situation has sparked controversy among scientists, many of whom have urged the PTO not to grant broad patents at this early stage to applicants who have neither characterized the genes nor determined their functions and specific uses.

Genes and other biological resources have been patentable since the landmark 1980 U.S. Supreme Court decision in *Diamond v Chakrabarty* that granted a patent for an oil-dissolving microbe. Patents give owners exclusive rights to their inventions or ideas for 20 years from the filing date. The rationale is to allow inventors time to recoup their investment costs in exchange for a public description of their knowledge, thereby revealing technical advances to competitors and the general public and avoiding duplicated efforts. Biological inventions are patentable if they meet the standard requirements for all patents: they must be novel, useful, not obvious, and described sufficiently for others to reproduce.

A single gene may be patented, in principle, by different scientists or companies. One concern is that such "patent stacking" may discourage product development because royalties might be owed to all patent owners. Additionally, because applications remain secret, companies may work on developing a product, only to find that "submarine patents" already have been granted, leading to unexpected licensing costs and possible infringement penalties.

Some past controversies have centered around the "utility" requirement. Some fear the large-scale patenting of gene fragments by biotechnology companies who are unaware of their functions but would stake a claim to all future

(see Patents, p. 4)

More patenting information:

- ***www.ornl.gov/hgmis/elsi/patents.html***

Human Genome Project FAQs

When is a Genome Completely Sequenced?

In December 1999, the 56-Mb sequence of human chromosome 22 was declared essentially complete, yet only 33.5 Mb were sequenced. In early spring of 2000, the fruit fly *Drosophila's* 180-Mb genome also was announced as completed, although just 120 Mb were characterized. What's the deal?

Animal genomes have large DNA regions that currently cannot be cloned or assembled. In the human genome sequence, these regions include telomeres and centromeres (chromosome tips and centers), as well as many chromosomal areas packed with other types of sequence repeats.

Most unsequenceable areas contain heterochromatic DNA, which has few genes and many repeated regions that are difficult to maintain as clones for DNA sequencing. HGP scientists strive to sequence the entire euchromatic DNA, which generally is defined as gene-rich areas (including both exons and introns) that are translated into RNA during gene expression. In the case of human chromosome 22, the sequenced 60% represents 97% of euchromatic DNA. Similarly, nearly all the euchromatic regions were sequenced for *Drosophila*.

Although the HGP goal is to have complete strings of sequence for each chromosome from tip to tip, obtaining this high level of resolution presents a great challenge.

Whose Genomes?

All humans share the same basic set of genes and genomic regulatory regions that control the development and maintenance of biological structures and processes. Therefore, the human reference sequence will not, and does not need to, represent an exact match for any one person's genome.

Investigators are using DNA from donors representing widely diverse populations. For example, HGP researchers collected samples of blood (female) or sperm (male) from a large number of people; only a few samples were processed, with source names protected so neither donors nor scientists would know whose genomes were being sequenced. The private company Celera Genomics collected samples from five individuals who identified themselves as Hispanic, Asian, Caucasian, and African-American.

In addition to generating the reference sequence, another important HGP goal is to identify many of the small DNA regions that vary among individuals and could underlie disease susceptibility and drug responsiveness. The most common variations are called SNPs (single nucleotide polymorphisms). The DNA resources used for these studies came from 24 anonymous donors of European, African, American (north, central, south), and Asian ancestry.

Although the sequence information will come from the DNA of many persons, it will be applicable to everyone.

Draft vs Finished Sequence

In generating the draft sequence, scientists determined the order of base pairs in each chromosomal area at least 4 to 5 times (4× to 5×) to ensure data accuracy and to help with reassembling DNA fragments in their original order. This repeated sequencing is known as genome "depth of coverage." Draft sequence data are mostly in the form of 10,000 bp–sized fragments whose approximate chromosomal locations are known.

To generate finished high-quality sequence, additional sequencing is needed to close gaps, reduce ambiguities, and allow for only a single error every 10,000 bases, the agreed-upon standard for HGP finished sequence. Investigators believe that a high-quality sequence is critical for recognizing regulatory components of genes that are very important in understanding human biology and such disorders as heart disease, cancer, and diabetes. The finished version will provide an estimated 8× to 9× coverage of each chromosome. Thus far, finished sequences have been generated for only two human chromosomes–21 and 22.

Why DOE?

DOE's role in the HGP arose from the historic congressional mandate of its predecessor agencies (the Atomic Energy Commission and the Energy Research and Development Administration) to study the genetic and health effects of radiation and chemical by-products of energy production. From this work the recognition grew that the best way to learn about these effects was to study DNA directly.

What's the Next Step?

Building on data produced in the HGP, a new DOE research program, Genomes to Life, is aimed at accelerating the understanding of living systems (*DOEGenomesToLife.org*). ♦

Patents *(from p. 3)*

discoveries on those genes (sometimes called "reach-through patents").

In December 1999, the PTO published revised interim guidelines clarifying the utility requirement for patent claims on genomic and other biotechnological inventions. The interim guidelines called for "specific and substantial utility that is credible," but some still felt they were not stringent enough. Public comments were posted to the PTO Web site (*www.uspto.gov*; click on "Site Index," then "P" for Public Comments).

On January 5, 2001, PTO responded to public comments and issued final guidelines that were largely unchanged (*www.uspto.gov/web/offices/com/sol/notices/utilexmguide.pdf*). *[Denise Casey, HGMIS]* ♦

Human Genome news

Human Genome Management Information System (HGMIS)
865/576-6669, Fax: /574-9888
mansfieldbk@ornl.gov; www.ornl.gov/hgmis

This newsletter is prepared at the request of the DOE Office of Biological and Environmental Research by the Life Sciences Division at Oak Ridge National Laboratory, which is managed by UT-Battelle, LLC, under contract AC05-00OR22725.