In the News

Human Genome Working Draft: First-Edition Travel Guides

In February, scientists from the public Human Genome Project and the private company Celera Genomics published the long-awaited details of the working-draft DNA sequence achieved less than a year before. Although the draft is filled with mysteries, the first panoramic view of the human genetic landscape has revealed a wealth of information and some early surprises. Papers describing research observations in the journals *Nature* (Feb. 15) and *Science* (Feb. 16) are freely accessible via the Web (*www.ornl.gov/hgmis/project/journals/journals.html*).

Although clearly not a Holy Grail or Rosetta Stone for deciphering all of biology—two early metaphors commonly used to describe the coveted prize—the sequence is a magnificent and unprecedented resource that will serve as a basis for research and discovery throughout this century and beyond. It will have diverse practical applications and a profound impact upon how we view ourselves and our place in the tapestry of life around us.

One insight already gleaned from the sequence is that, even on the molecular level, we are more than the sum of our 35,000 or so genes. Surprisingly, this newly estimated number of genes is only one-third as great as previously thought and is only twice as many as those of a tiny transparent worm, although the numbers may be revised as more computational and experimental analyses are performed. At once humbled and intrigued by this finding, scientists suggest that the genetic key to human complexity lies not in the number of genes but in how gene parts are used to build different products in a process called alternative splicing. Other sources of added complexity are the thousands of post-translational chemical modifications made to proteins and the repertoire of regulatory mechanisms controlling these processes.

The draft encompasses 90% of the human genome's euchromatic portion, which contains the most genes. In constructing the working draft, the 16 genome sequencing centers produced over 22.1 billion bases of raw sequence data, comprising overlapping fragments totaling 3.9 billion bases and providing sevenfold coverage (sequenced seven times) of the human genome. Over 30% is high-quality, finished sequence, with eight- to tenfold coverage, 99.99% accuracy, and few gaps. All data are freely available via the Web (*www.ornl.gov/hgmis/project/journals/sequencesites.html*).

The entire working draft will be finished to high quality by 2003. Coincidentally, that year also will be the 50th anniversary of Watson and Crick's publication of DNA structure that launched the era of molecular genetics (*www.nature.com/genomics/human/watson-crick*). Much will remain to be deciphered even then. Some highlights from *Nature*, *Science*, and The Wellcome Trust follow (*www.ornl.gov/hgmis/project/journals/insights.html*).

What Does the Draft Human Genome Sequence Tell Us?

By the Numbers

- The human genome contains 3164.7 million chemical nucleotide bases (A, C, T, and G).
- The average gene consists of 3000 bases, but sizes vary greatly, with the largest known human gene being dystrophin at 2.4 million bases.
- The total number of genes is estimated at 30,000 to 35,000—much lower than previous estimates of 80,000 to 140,000 that had been based on extrapolations from gene-rich areas as opposed to a composite of gene-rich and gene-poor areas.
- Almost all (99.9%) nucleotide bases are exactly the same in all people.
- The functions are unknown for over 50% of discovered genes.

The Wheat from the Chaff

- Less than 2% of the genome codes for proteins.
- Repeated sequences that do not code for proteins ("junk DNA") make up at least 50% of the human genome.
- Repetitive sequences are thought to have no direct functions, but they shed light on chromosome structure and dynamics. Over time, these repeats reshape the genome by rearranging it, creating entirely new genes, and modifying and reshuffling existing genes.
- During the past 50 million years, a dramatic decrease seems to have occurred in the rate of accumulation of repeats in the human genome.

How It's Arranged

- The human genome's gene-dense "urban centers" are predominantly composed of the DNA building blocks G and C.
- In contrast, the gene-poor "deserts" are rich in the DNA building blocks A and T. GC- and AT-rich regions usually can be seen through a microscope as light and dark bands on chromosomes.
- Genes appear to be concentrated in random areas along the genome, with vast expanses of noncoding DNA between.
- Stretches of up to 30,000 C and G bases repeating over and over often occur adjacent to gene-rich areas, forming a barrier between the genes and the "junk DNA." These

CpG islands are believed to help regulate gene activity.

- Chromosome 1 has the most genes (2968), and the Y chromosome has the fewest (231).

How the Human Compares with Other Organisms

- Unlike the human's seemingly random distribution of gene-rich areas, many other organisms' genomes are more uniform, with genes evenly spaced throughout.
- Humans have on average three times as many kinds of proteins as the fly or worm because of mRNA transcript "alternative splicing" and chemical modifications to the proteins. This process can yield different protein products from the same gene.
- Humans share most of the same protein families with worms, flies, and plants, but the number of gene family members has expanded in humans, especially in proteins involved in development and immunity.
- The human genome has a much greater portion (50%) of repeat sequences than the mustard weed (11%), the worm (7%), and the fly (3%).
- Although humans appear to have stopped accumulating repeated DNA over 50 million years ago, there seems to be no such decline in rodents. This may account for some of the fundamental differences between hominids and rodents, although gene estimates are similar in these species. Scientists have proposed many theories to explain evolutionary contrasts between humans and other organisms, including those of life span, litter sizes, inbreeding, and genetic drift.

Variations and Mutations

- Scientists have identified about 1.4 million locations where single-base DNA differences (SNPs) occur in humans. This information promises to revolutionize the processes of finding chromosomal locations for disease-associated sequences and tracing human history.
- The ratio of germline (sperm or egg cell) mutations is 2:1 in males vs females. Researchers point to several reasons for the higher mutation rate in the male germline, including the greater number of cell divisions required for sperm formation than for eggs.

HGP Sequenced Genomes

Organism	Size (Mb)	Yr. Seq.	No. of Genes*	Gene Density**
Saccharomyces cerevisiae yeast (eukaryote)	12.1	1996	6034[1]	483
Escherichia coli bacterium (prokaryote)	4.6	1997	4200[2]	932
Caenorhabditis elegans roundworm (eukaryote)	97	1998	19,099[1]	197
Arabidopsis thaliana plant (eukaryote)	100	2000	25,000[1]	221
Drosophila melanogaster fruit fly (eukaryote)	180	2000	13,061[1]	117
Homo sapiens human (eukaryote)	3000	2000 draft	35,000–45,000[1]	12

Sources

1. *Nature* **409**, 819 (Feb. 15, 2001) (*www.nature.com/nature/journal/v409/n6822/fig_tab/409818a0_F1.html*)

2. Entrez Genomes (*www.ncbi.nlm.nih.gov/PMGifs/Genomes/micr.html*)

*Gene predictions are made by computational algorithms based on recognition of gene-sequence features and similarities to known genes. Current gene estimates await further confirmation, including characterization of their protein products and functions.

**Gene density = Number of genes per million sequenced DNA bases.

Applications, Future Challenges

Deriving meaningful knowledge from the DNA sequence will define research through the coming decades to inform our understanding of biological systems. This enormous task will require the expertise and creativity of tens of thousands of scientists from varied disciplines in both the public and private sectors worldwide.

The draft sequence already is having an impact on finding genes associated with disease. Over 30 genes have been pinpointed and associated with breast cancer, muscle disease, deafness, and blindness. Additionally, finding the DNA sequences underlying such common diseases as cardiovascular disease, diabetes, arthritis, and cancers is being aided by the human variation maps (SNPs) generated in the HGP in cooperation with the private sector. These genes and SNPs provide focused targets for the development of effective new therapies.

One of the greatest impacts of having the sequence may well be in enabling an entirely new approach to biological research. In the past, researchers studied one or a few genes at a time. With whole-genome sequences and new high-throughput technologies, they can approach questions systematically and on a grand scale. They can study all the genes in a genome, for example, or all the transcripts in a particular tissue or organ or tumor, or how tens of thousands of genes and proteins work together in interconnected networks to orchestrate the chemistry of life.

Post-sequencing projects are well under way worldwide (see related articles, pp. 1 and 7). These explorations will result in a more comprehensive, new, and profound understanding of complex living systems, with applications to human health, energy, global climate change, and environmental cleanup, among others. *[Denise Casey, HGMIS]* ◊

Take a Genome Tour

National Center for Biotechnology Information (NCBI) tutorial on how to use the publicly available DNA sequence data and analysis tools:
www.ncbi.nlm.nih.gov/Tour

NCBI Human Genome Map Viewer:
www.ncbi.nlm.nih.gov/cgi-bin/Entrez/hum_srch

Guide to online information resources:
www.ncbi.nlm.nih.gov/genome/guide/human