A few highlights from the HGP databanks:

- *The human genome contains 3 billion chemical nucleotide bases (A, C, T, and G). Compare this to the laboratory mouse (2.6 billion), the fruitfly (137 million), yeast (12.1 million), and the bacterium E. coli (4.6 million).*
- *The average gene consists of 3,000 bases, but sizes vary greatly, with the largest known human gene being dystrophin at 2.4 million bases.*
- *The functions are unknown for more than 50 percent of discovered genes.*
- *The human genome sequence is almost (greater than 99 percent) the same in all people.*
- *About 2 percent of the genome encodes instructions for the synthesis of proteins.*
- *More than 40 percent of the predicted human proteins share similarity with fruitfly or worm proteins.*
- *Genes appear to be concentrated in random areas along the genome, with vast expanses of DNA between.*
- *Chromosome 1 (the largest human chromosome) has the most genes (2,968), and the Y chromosome has the fewest (231).*

Want more information:

The Human Genome Project:
www.genome.gov/10001772

The International HapMap:
www.hapmap.org

ENCODE:
www.genome.gov/10005107

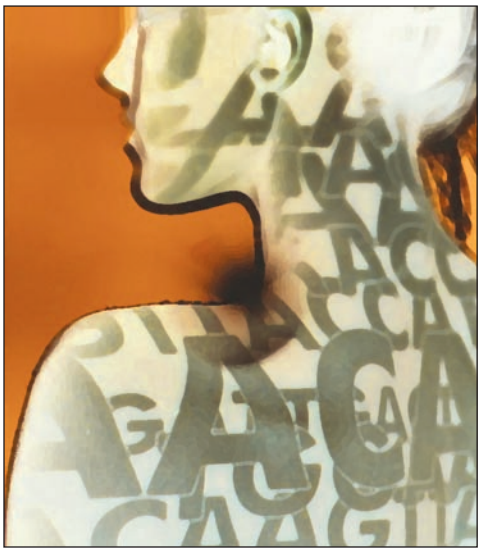
A quick guide to sequenced genomes from the Genome News Network:
www.genomenewsnetwork.org/resources/sequenced_genomes/genome_guide_p1.shtml

BIOTECH 101

The Human Genome Project: looking back, looking ahead

Five years ago this spring, the "completion" of the Human Genome Project (HGP) was announced with much fanfare. This 13-year collaboration identified the sequence of the 3 billion chemical bases of information that reside on 46 chromosomes in nearly every cell in our body. The published DNA sequence was akin to an operations manual or book of recipes, identifying the genetic instructions for how cells build, operate, maintain and reproduce, all while responding to varying conditions from the surrounding environment. (See the listing, left, for some of the HGP's findings.) The finished sequence was hailed as biology's "moonshot" – accomplished by the collaborative efforts of over 2,000 scientists spread across six countries. Coming in under budget and two years ahead of schedule, it was truly an amazing effort.

While the completion of the HGP may have felt like the end of an era, in reality it was only the beginning. Scientists had very little knowledge of how cells utilized the information found in each genetic recipe to function and interact with each other and the environment. Furthermore, they didn't have a clear understanding of how genes keep us healthy or predispose us to disease. A representative genome had been sequenced, but how many differences would be found if peoples from around the world were compared? How did the human sequence compare to those of other organisms? Sequencing the human genome raised more questions than it answered.



Flash forward to 2008 and while questions still abound, scientists are beginning to compile an impressive array of results. In this issue of Biotech 101, we'll examine the impact of the HGP, highlighting three research areas. Just like the HGP, all the information generated from these resulting fields are freely accessible by scientists and the public around the world – see the "Want more information" section to link directly to the research.

Genetic Variation

Any two humans are identical at the 99.9 percent level. While there are many more similarities than differences, slight variations in our DNA can have a major impact on whether or not we develop a particular disease, how we respond to an infection and which drugs are most effective. In order to understand the effects of this genetic variation, we first need to identify and catalog the similarities and differences found across human populations. The International HapMap Project was created in 2003 to compare the genetic sequences of different individuals. Internationally funded, the project is a collaboration between scientists from Japan, Canada, the United

Kingdom, China, Nigeria and the United States. The HapMap describes where in DNA the variants are found and how they are distributed within and across world populations. The project does not connect the variation to a specific illness, but rather provides the raw information that researchers can use to link genetic variation to disease risk. Although only a few years old, the information gleaned by the HapMap is beginning to yield results. Working with the HapMap's catalog of variation, researchers have identified genetic variation that contributes to risk for a number of diseases, including type 2 diabetes, Parkinson's disease, heart disorders, obesity, Crohn's disease, prostate cancer and age-related macular degeneration.

Controls and Signals

ENCODE, the **ENC**yclopedia **O**f **D**NA **E**lements, was launched in 2003 to develop efficient ways of identifying and precisely locating all of the protein-coding genes, non-protein-coding genes and other functional components in the human genome. The goal of ENCODE is to develop the parts list of biologically functional elements and subsequently determine the signals that activate or silence each of the parts, identifying how such signals interact with each other and the environment.

Last summer, the ENCODE team published preliminary findings based on a pilot set of data. Suprisingly, the organization and function of the genome appears to be far more complicated than most had suspected. The ENCODE data indicate that the majority of DNA in the human genome has some sort of functional role. This challenges the long-standing view that the human genome consists of a relatively small set of functional elements (the genes) along with a vast amount of so-called "junk" DNA that is not biologically active. The new data indicate the genome contains very few unused sequences and, in fact, is a complex, interwoven network.

One member of the multi-disciplinary ENCODE team is a familiar face at HudsonAlpha, Dr. Richard Myers, the institute's director. Dr. Myers' lab is identifying sequences in the genome where proteins or enzymes bind and activate DNA sequences. He is also testing a new approach to determine the methylation status of specific DNA regions. Methylation refers to a specific chemical modification of DNA which can silence or reduce the activity of the affected DNA region. (Refer to the fall 2007 Biotech 101 article on epigenetics for more details about methylation and its role in gene activity.)

Comparative Genomics

Comparative genomics examines and compares the genome sequences of different species - human, rat and a wide variety of other creatures from roundworms to roosters. Since the mid 1990s, over 180 organisms have had their genomes sequenced. By comparing the human reference sequence with genomes of other organisms, researchers can identify areas of similarity and difference and better understand a gene's structure and function. For example, researchers have found that two-thirds of human genes known to be involved in cancer have counterparts in the fruit fly. Research using organisms such as flies, worms and mice offers a cost-effective way to examine the functions of these similar genes. Increased understanding of genes' connections to diseases supports opportunities for potential treatments and pharmaceutical interventions.

Comparative genomics also provides a powerful tool for studying evolutionary changes among organisms, helping to identify genes that are conserved among species (which may represent critical biological functions across the tree of life), as well as genes that give each organism its unique characteristics.

The Future

The completion of the Human Genome Project has allowed researchers to initiate the shift from sequence determination to functional understanding. In many ways, the real work has just begun on this journey into the genome era. Given the breathtaking pace of discovery from projects like HapMap and ENCODE and the wide array of genomes from other organisms now available for study, it will be exciting to watch as the impact expands across biology, health and society. ■

– Dr. Neil Lamb
director of educational outreach
HudsonAlpha Institute for Biotechnology