International Human Genome Sequencing Consortium Describes Finished Human Genome Sequence Researchers Trim Count of Human Genes to 20,000-25,000

BETHESDA, Md., Wed., Oct. 20, 2004 - 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 published its scientific description of the finished human genome sequence, reducing the estimated number of human protein-coding genes from 35,000 to only 20,000-25,000, a surprisingly low number for our species.

The paper appears in the Oct. 21 issue of the journal *Nature*. In the paper, researchers describe the final product of the Human Genome Project, which was the 13-year effort to read the information encoded in the human chromosomes that reached its culmination in 2003. The *Nature* publication provides rigorous scientific evidence that the genome sequence produced by the Human Genome Project has both the high coverage and accuracy needed to perform sensitive analyses, such as focusing on the number of genes, the segmental duplications involved in disease and the "birth" and "death" of genes over the course of evolution.

"Only a decade ago, most scientists thought humans had about 100,000 genes. When we analyzed the working draft of the human genome sequence three years ago, we estimated there were about 30,000 to 35,000 genes, which surprised many. This new analysis reduces that number even further and provides us with the clearest picture yet of our genome," said NHGRI Director Francis S. Collins, M.D., Ph.D. "The availability of the highly accurate human genome sequence in free public databases enables researchers around the world to conduct even more precise studies of our genetic instruction book and how it influences health and disease."

One of the central goals of the effort to analyze the human genome is the identification of all genes, which are generally defined as stretches of DNA that code for particular proteins. According to the new findings, researchers have confirmed the existence of 19,599 protein-coding genes in the human genome and identified another 2,188 DNA segments that are predicted to be protein-coding genes.

"The analysis found that some of the earlier gene models were erroneous due to defects in the unfinished, draft sequence of the human genome," said Jane Rogers, Ph.D., head of sequencing at the Wellcome Trust Sanger Institute in Hinxton, England. "The task of identifying genes remains challenging, but has been greatly assisted by the finished human genome sequence, as well as by the availability of genome sequences from other organisms, better computational models and other improved resources."

The Nature paper also provides the scientific community with a peer-reviewed description of the finishing process, and an assessment of the quality of the finished human genome sequence, which was deposited into public databases in April 2003. The assessment confirms that the finished sequence now covers more than 99 percent of the euchromatic (or gene-containing) portion of the human genome and was sequenced to an accuracy of 99.999 percent, which translates to an error rate of only 1 base per 100,000 base pairs - 10 times more accurate than the original goal.

The contiguity of the sequence is also massively improved. The average DNA letter now sits on a stretch of 38.5 million base pairs of uninterrupted, high-quality sequence - about 475 times longer than the 81,500 base-pair stretch that was available in the working draft. Access to uninterrupted stretches of sequenced DNA can greatly assist researchers hunting for genes and the neighboring DNA sequences that may regulate their activity, dramatically cutting the effort and expense required to find regions of the human genome that may contain small and often rare variants involved in disease.

"Finished" doesn't mean that the human genome sequence is perfect. There still remain 341 gaps in the finished human genome sequence, in contrast to the 150,000 gaps in the working draft announced in June 2000. The technology now available cannot readily close these recalcitrant gaps in the human genome sequence. Closing those gaps will require more research and new technologies, rather than industrial-scale efforts like those employed by the Human Genome Project.

"The human genome sequence far exceeds our expectations in terms of accuracy, completeness and continuity. It reflects the dedication of hundreds of scientists working together toward a common goal - creating a solid foundation for biomedicine in the 21st century," said Eric Lander, Ph.D., director of the Broad Institute of MIT and Harvard in Cambridge, Mass.

In addition to reducing the count of human genes, scientists reported that the improved quality of the finished human genome sequence, compared with earlier drafts, provides a much clearer picture of certain phenomena such as duplication of DNA segments and the birth and death of genes. Segmental duplications are large, almost identical copies of DNA, which are present in at least two locations in the human genome. A number of human diseases are known to be associated with mutations in segmentally duplicated regions, including Williams syndrome, Charcot-Marie-Tooth and DiGeorge syndrome.

"Segmental duplications were almost impossible to study in the draft sequence. Now, through the unstinting efforts of groups around the world, this important and rapidly evolving part of our genome is open for scientific exploration," said Robert H. Waterston, M.D., Ph.D., former director of the Genome Sequencing Center at Washington University in St. Louis and now chair of the Department of Genome Sciences at the University of Washington in Seattle. Segmental duplications cover 5.3 percent of the human genome, significantly more than in the rat genome, which has about 3 percent, or the mouse genome, which has between 1 and 2 percent.

Segmental duplications provide a window into understanding how our genome evolved and is still changing. The high proportion of segmental duplication in the human genome shows our genetic material has undergone rapid functional innovation and structural change during the last 40 million years, presumably contributing to unique characteristics that separate us from our non-human primate ancestors.

The consortium's analysis found the distribution of segmental duplications varies widely across human chromosomes. The Y chromosome is the most extreme case, with segmental duplications occurring along more than 25 percent of its length. Some segmental duplications tend to be clustered near the middle (centromeres) and ends (telomeres) of each chromosome, where, researchers postulate, they may be used by the genome as an evolutionary laboratory for creating genes with new functions.

The accuracy of the finished human genome sequence produced by the Human Genome Project has also given scientists some initial insights into the birth and death of genes in the human genome. Scientists have identified more than 1,000 new genes that arose in the human genome after our divergence with rodents some 75 million years ago. Most of these arose through recent gene duplications and are involved with immune, olfactory and reproductive functions. For example, there are two families of genes recently duplicated in the human genome that encode sets of proteins (pregnancy-specific beta-1 glycoprotein and choriogonadotropin beta proteins) that may be involved in the extended period of pregnancy unique to humans.

Additionally, researchers used the finished human genome to identify and characterize 33 nearly intact genes that have recently acquired one or more mutations, causing them to stop functioning, or "die." Scientists pinpointed these non-functioning genes, referred to as pseudogenes, in the human genome by aligning them with the mouse and rat genomes, in which the corresponding genes have maintained their functionality. Interestingly, researchers determined that 10 of these pseudogenes in the human genome sequence appear to have coded for proteins involved in olfactory reception, which helps to explain why humans have fewer functional olfactory receptors and, consequently, a poorer sense of smell than rodents. The molecular biology of the sense of smell was just recognized by the awarding of a Nobel Prize in Physiology or Medicine to Richard Axel and Linda B. Buck.

Next, the researchers aligned the 33 pseudogenes with the draft sequence of the chimpanzee genome to determine whether they were still functional before Homo sapiens' divergence from great apes about 5 million years ago. The analysis revealed that 27 of the pseudogenes were non-functional in both humans and chimps. However, five of the genes that were inactive in humans were found to be still functional in chimpanzees. "The identification of these pseudogenes and their functional counterparts in chimpanzee provides fertile ground for future research projects," said Richard Gibbs, Ph.D., director of Baylor College of Medicine's Human Genome Sequencing Center in Houston, which currently is sequencing the genome of another non-human primate, the rhesus macaque (Macaca mulatta).

More than 2,800 researchers who took part in the International Human Genome Sequencing Consortium share authorship on today's Nature paper, which expands upon the group's initial analysis published in Feb. 2001. Even more detailed annotations and analyses have already been published for chromosomes 5, 6, 7, 9, 10, 13, 14, 19, 20, 21, 22 and Y. Publications describing the remaining 12 chromosomes are forthcoming.

The finished human genome sequence and its annotations can be accessed through the following public genome browsers:

- GenBank at NIH's National Center for Biotechnology Information (NCBI)
- UCSC Genome Browser at the University of California at Santa Cruz
- Ensembl Genome Browser at the Wellcome Trust Sanger Institute and the EMBL-European Bioinformatics Institute
- DNA Data Bank of Japan
- EMBL-Bank at the European Molecular Biology Laboratory's Nucleotide Sequence Database

The International Human Genome Sequencing Consortium includes scientists at 20 institutions located in France, Germany, Japan, China, Great Britain and the United States. The five largest sequencing centers are located at: Baylor College of Medicine; the Broad Institute of MIT and Harvard; DOE's Joint Genome Institute, Walnut Creek, Calif.; Washington University School of Medicine; and the Wellcome Trust Sanger Institute.

For more information, contact:

Geoff Spencer, NIH NHGRI, 301/402-0911, spencerg@mail.nih.gov
Jeff Sherwood, Department of Energy Human Genome Program, 202-586-5806