EMBARGOED UNTIL 10 am EST , February 12, 2001

International Human Genome Sequencing Consortium Publishes Sequence and Analysis of the Human Genome

Washington, D.C., Feb. 12, 2001 – The Human Genome Project international consortium today announced ~~that it has published~~ a draft sequence and initial analysis of the human genome—the genetic blueprint for a human being. The paper will be published in the Feb. 15 issue of the journal *Nature*.

The draft sequence, which covers more than 90 percent of the human genome, represents the exact order of DNA's four chemical bases, commonly abbreviated A, T, C and G, arrayed along the human chromosomes. This text underlies and shapes every human trait—from eye color and height, to health, diseases, and behavior.

The Consortium's initial analysis of this text describes some of the insights that can be gleaned from the genome and represents scientists' first ~~glimpse~~ of the human DNA terrain— its quirks and curiosities, as well as the extraordinary trove of information about human development, physiology, medicine, and evolution.

The results reported in *Nature* paper represents ~~the next milestone~~ since the consortium's announcement on June 26 that it had ~~completed a first assembly of the human genome. The first assembly announced in June was analogous to having a collection of all the letters that make up~~ 90 percent of the text for the "Book of Life." This week's paper represents a compilation of these ~~scrambled~~ letters into the first draft of a readable text, ~~and an in depth analysis of the Script, which reveals many surprises.~~

There are ~~a few~~ small gaps still remaining to be filled in this text, but scientists are already getting a ~~gist~~ of what the genome topography looks like and the stories it has to tell. Below are highlights.

- The mammalian genome has a striking geography. It looks more like the ~~Rocky Mountains~~, with jagged peaks and ~~deep ravines~~, and less like the cornfields of Iowa. ~~The genes live on the mountain tops and the 'junk DNA' live in the less desirable valleys~~. The genome is filled with crowded urban areas, fancy neighborhoods, and deserted ghost towns—each telling a unique story

about the history of our species and what makes us tick. This topography is in striking contrast to the genomes of many other organisms, such as the mustard weed, the worm, and the fly. Their genomes, it seems, are ~~flatter than the~~ more like the midwestern plains—with a more evenly distributed population density of genes. (See Vignette 1).

and gene detection methods need to be improved

- Although small gaps in the human genome sequence must be filled before scientists can arrive at an exact number, scientists can now say that humans have some 30-35,000 genes in their genomes. These new numbers indicate that humans have only twice as many genes as the worm or the fly! How can human complexity be explained by a genome with such paucity in genes? It turns out humans are very thrifty with their genes, able to do more with what they have than other species. Instead of producing only one protein per gene, human genes may be able to produce five different proteins. (See Vignette 2)

average three

- The full set of proteins (the proteome) encoded by the human genome is more complex than those of invertebrates, but that's largely because vertebrates appear to have rearranged old protein domains into a richer collection of new architectures. And not because we have made new innovations in our genome. ← Quite the opposite in fact: humans have achieved innovations by rearranging and expanding tried-and-true strategies from other species. (See Vignette 3)

- Scientists have identified more than two hundred genes in the human genome whose closest relatives are in bacteria. Analogous genes are not found in invertebrates, such as the worm, fly, and yeast, suggesting that these genes were acquired at a more recent evolutionary past, perhaps after the birth of vertebrates. Most probably, infections led to a transfer of DNA from bacteria to the chromosomes of a human ancestor. Scientists didn't find any single bacterial source for the transferred genes, indicating that several independent gene transfers from different bacteria occurred. (See Vignette 9)

or viral agents

These genes help us!

- Our junk DNA represents a rich fossil record of clues to our evolutionary past. It is possible to date groups of repeats to when they were "born" during evolution and to follow their fates in different regions of the genome or in different species. The HGP scientists used 3 million such elements as dating tools. Based on such "DNA dating," scientists can build family trees of the repeats showing exactly where they came from and when. These repeats have reshaped the genome by rearranging it, creating entirely new genes, and modifying and reshuffling existing genes. (See Vignette 4)

Can see back 500×10^6 years

- We have more repeats, or junk DNA, in our genomes—50 percent—than the mustard weed (11 percent), the worm (7 percent) or the fly (3 percent). Also, shockingly, there seems to have been a dramatic decrease in repeats in the

human genome over the past 50 million years. It's as if we decided at that point to stop collecting junk. In contrast, there seems to be no such decline in repeats in rodents. (See Vignette 6)

with few genes and

- Most repeat elements—second class citizens—land up in less desirable neighborhoods in the genome—regions that are AT rich and GC poor. ~~This was true for the LINE elements, the most privileged of the repeats, the least junky of the junk~~. But mysteriously, SINE elements, ~~the freeloaders that piggyback on LINEs~~, seem to have landed in the fancy, GC rich neighborhoods of the genome. SINE elements may have a bad rap for parasite-like behavior. In fact ~~they~~ are selected because they are doing us good—they are symbiots that earn their keep in the genome. (See Vignette 7)

This part of the junk isn't junk after all.

this new information suggests

- By dating the 3 million repeat elements, and examining the pattern of interspersed repeats on Y chromosome, scientists estimated the relative mutation rates in the X and the Y and in the male and female germ lines. They found that the ratio of mutations in males versus female is 2.1.—confirming an estimate by David Page and his colleagues at the Whitehead Institute. (See Vignette 8)

so what

- In a companion volume to the book of life, scientists have created the largest publicly available catalogue of single letter differences (SNPs)—1.4 million SNPs—with their exact location in the human genome. The SNP map promises to revolutionize both mapping diseases and tracing human history. (See Vignette 10)

The sequence information from the public project has been immediately and freely released to the world, with no restrictions on its 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. Already, many tens of thousands of genes have been identified from the genome sequence.

30 disease genes

The scientific work reported here will serve as a basis for research and discovery in the coming decades. Such research will have profound long-term consequences for medicine and help elucidate the underlying molecular mechanisms of disease. This will in turn allow researchers to design better drugs and therapies for many illnesses.

"But the science is only part of the challenge," write the authors of the Nature paper. "We must also involve society at large in the work ahead. We must set realistic expectations that the most important benefits will not be reaped overnight. Moreover, understanding and wisdom will be required to ensure that they are implemented broadly and equitably."

"We are standing at an extraordinary moment in scientific history. It's as though we have climbed to the top of the Himalayas, and we can for the first time see the breathtaking vista of the human genome," says Eric Lander, Director of the Whitehead Institute Center for Genome Research. "We can see the terrain in a new light, but we still have many mountains to climb and valleys to cross before we will intimately understand all the secrets that the genome has to tell us."

"This remarkable achievement is a clear testament to the hard work of the hundreds of scientists in the sixteen genome centers that make up the Human Genome Project Consortium," said Francis Collins, Director of the National Human Genome Research Institute. "The winners ~~are the~~ scientists ~~they~~ have proved to the world that they can work together toward a common human good. For, with the human genome sequence in hand, we can begin to build the tools we need to conquer the host of illnesses ~~in this world."~~

The consortium's ultimate goal is to produce a completely "finished" sequence—with no gaps and 99.99 percent accuracy. The target date for this ultimate goal had been 2003, but today's results mean that the goal of a final, stand-the-test-of-time sequence will be met considerably ahead of schedule.

Next phase. The next phase of the Human Genome Project will focus on its ultimate goal of converting the draft and near-finished sequences to a finished form, by filling the gaps in the sequence and by increasing the overall sequence accuracy to 99.99 percent. Although the near-finished version is adequate for the most biomedical research, the HGP has made a commitment to filling all gaps and resolving all ambiguities in the sequence.

Production of genome sequence has skyrocketed over the past year, with more than 90 percent of the sequence having been produced in the past fifteen months alone. Because of this increased capacity, the next phase is expected to move ~~much~~ even more rapidly than previously expected.

The HGP also plans to sequence the genomes of many species, for comparing genomes across species will provide researchers key tools to understanding the essential elements that evolution has designated as important to survival. This information will in turn translate to practical knowledge toward developing better therapies in the future.

The HGP sequencing consortium used a biocluster provided by Compaq Computer Corporation's High Performance Technical Computing group. The

27 node configuration of AlphaServer ES40s, containing 108 CPUs provided one terabyte of secondary storage and assisted annotation and analysis.

In a related announcement today, the biotech firm Celera Genomics announced that it had also published its sequence in the journal *Science.*

Background

Sequencing, which is determining the exact order of DNA's four chemical bases, commonly abbreviated A, T, C and G, has been expedited in the Human Genome Project by technological advances in deciphering DNA and the collaborative nature of the effort, which includes about 1,000 scientists worldwide working together effectively.

The Human Genome Sequencing Project aims to determine the sequence of the *euchromatic* portion of human genome. The *euchromatic* portion excludes certain regions consisting of long stretches of highly repetitive DNA that encode little genetic information. Such regions are said to be *heterochromatic.* (Genomes contain certain portions consisting of long stretches of highly repetitive DNA (for example, the center of chromosomes, called centromeres, consists of heterochromatic DNA.)

The international Human Genome Sequencing consortium includes scientists at 16 institutions in France, Germany, Japan, China, Great Britain and the United States. The five largest centers are located at: Baylor College of Medicine, Houston, Texas; Joint Genome Institute in Walnut Creek, CA; Sanger Centre near Cambridge, England; Washington University School of Medicine, St. Louis; and Whitehead Institute, Cambridge, Massachusetts. The attached table provides more detail about the 16 centers and their individual contributions to the Human Genome Project.

The project is funded by grants from government agencies and public charities in the various countries. These include the National Human Genome Research Institute at the US National Institutes of Health, the Wellcome Trust in England, and the US Department of Energy.

The total cost for Phase One ("working draft") is approximately $300 million worldwide, with roughly half ($150 million) being funded by the US National Institutes of Health.

The Human Genome Project is sometimes reported as having a cost of $3

billion. However, this figure refers to the total projected funding over a 15-year period (1990-2005) for a wide range of scientific activities related to genomics. These include studies of human diseases, experimental organisms (such as bacteria, yeast, worms, flies and mice), development of new technologies for biological and medical research, computational methods to analyze genomes, and ethical, legal and social issues related to genetics. Human Genome Sequencing represents only a small fraction of the overall fifteen-year budget.

###

The international Human Genome Sequencing Consortium includes scientists at institutions in France, Germany, Japan, China, Great Britain, Canada and the United States.

The sixteen 20 institutions that form the Human Genome Sequencing Consortium include: NO! This is the wrong #!!

1. Washington University School of Medicine Genome Sequencing Center, St. Louis, MO, USA
2. The Sanger Centre at the Wellcome Trust Genome Campus, Hinxton, UK
3. National Center for Biotechnology Information, NIH, Bethesda, MD, USA
4. Whitehead Institute for Biomedical Research, MIT, Cambridge, MA, USA
5. Albert Einstein College of Medicine, New York, NY, USA
6. Baylor College of Medicine Human Genome Sequencing Center, Houston, Texas, USA
7. Roswell Park Cancer Institute, Buffalo, NY, USA
8. University of Washington Multimegabase Sequencing Center, Seattle, WA, USA
9. Fred Hutchinson Cancer Research Institute, Seattle, WA, USA
10. Genoscope, Evry, France
11. U.S. Dept. of Energy Joint Genome Institute, Walnut Creek, CA, USA
12. Stanford Human Genome Center and Department of Genetics, Palo Alto, CA, USA
13. University of California, Santa Cruz, Santa Cruz, CA, USA
14. British Columbia Cancer Research Center, Vancouver BC, Canada
15. Department of Genome Analysis, Institute of Molecular Biotechnology, Jena, Germany
16. Department of Human Genetics and Pediatrics, University of California, Los Angeles, CA, USA
17. RIKEN Genomic Sciences Center, Saitama, Japan
18. Department of Molecular Biology, Keio University School of Medicine, Tokyo, Japan
19. Max-Planck-Institute for Molecular Genetics, Berlin, Germany