



URL of this page: <http://www.genome.gov/10005831>



NIH NEWS ADVISORY

National Institutes of Health

National Human Genome Research Institute

The Mouse Genome And The Measure of Man

December 2002

WASHINGTON, DC - The international Mouse Genome Sequencing Consortium today announced the publication of a high-quality draft sequence of the mouse genome - the genetic blueprint of a mouse - together with a comparative analysis of the mouse and human genomes describing insights gleaned from the two sequences. The paper appears in the Dec. 5 issue of the journal *Nature*.



The achievement represents a landmark advance for the Human Genome Project. It is the first time that scientists have compared and contrasted the contents of the human genome with that of another mammal. This milestone is all the more significant given that the laboratory mouse is the most important animal model and is widely used in the study of human diseases.

"Publishing the sequence in 2001 of the first mammalian genome - our own - was a remarkable and historical achievement. To sequence another mammalian genome in less than two years and to discover the treasure trove of information one can derive from a comparison of the two is beyond nearly anyone's dreams. It constitutes a tremendously exciting and defining moment for biomedical research," said Francis S. Collins, M.D., Ph.D., director of the National Human Genome Research Institute (NHGRI).

The mouse sequence provides scientists a powerful research tool to extract meaning from the human genome sequence, the "Book of Life" published in draft form last year. It allows them to recognize functionally important regions in the human genome by virtue of the fact that they are conserved through the 75 million years of evolution separating humans and mice. (See: [Background on Comparative Genomic Analysis](#))

"This is an extraordinary milestone. For the first time we have an opportunity to see ourselves in an evolutionary mirror," says Eric Lander, Ph.D., director of the Whitehead/MIT Center for Genome Research. "The mouse genome represents a very important chapter in evolution's lab notebook. Being able to read this notebook and compare genomic information across species allows us to glean important information about ourselves."

Because the mouse carries virtually the same set of genes as the human but can be used in laboratory research, this information will allow scientists to experimentally test and learn more about the function of

human genes, leading to better understanding of human disease and improved treatments and cures. (See: [Background on Mouse as a Model Organism](#))

"The mouse genome sequence will change the way research is done in this important experimental animal, just as genome sequences have opened new avenues of study for yeast, worms, and flies" says Robert Waterston, M.D., Ph.D., director of the Genome Sequencing Center at Washington University School of Medicine. "The genome sequence allows us to study genes and proteins in context and will spur the development of methods to study many genes in parallel. This more detailed molecular understanding of mouse biology will in turn produce new opportunities for understanding human disease and for devising effective therapies."

The draft sequence was assembled by the Mouse Genome Sequencing Consortium, an international team of scientists at the Whitehead Institute in Cambridge, Mass., Washington University in St. Louis, and the Wellcome Trust Sanger Institute and the European Bioinformatics Institute, in Hinxton, England. The project was funded in part by NHGRI of the U.S. National Institutes of Health (NIH) and the Wellcome Trust in the U.K. The sequencing centers were joined in the analysis effort by scientists from 27 institutions in six countries. These include computational biologists at University of California Santa Cruz in Santa Cruz, Calif., Pennsylvania State University in University Park, Penn., University of Oxford in Oxford, England, The Institute for Systems Biology in Seattle, Washington University School of Medicine in St. Louis, the University of Geneva in Switzerland and the Universitat Pompeu Fabra in Barcelona, Spain, among others. (See: [Nature](#) paper for a complete listing of authors and institutions)

"This is another stunning success for a publicly funded international consortium that is already acting to stimulate biomedical research," says Jane Rogers, Ph.D., head of Sequencing at The Wellcome Trust Sanger Institute. "Its value is predicated upon the quality of both the sequence and the tools to interpret it, all of which are freely available through Ensembl and similar browsers. Last week, as many people used the mouse sequence in Ensembl as used the human."

The sequence shows the order of the DNA chemical bases A, T, C, and G along the 20 chromosomes of a female mouse of the "Black 6" strain - the most commonly used mouse in biomedical research. It includes more than 96 percent of the mouse genome with long, continuous stretches of DNA sequence and represents a seven-fold coverage of the genome. This means that the location of every base, or DNA letter, in the mouse genome was determined an average of seven times, a frequency that ensures a high degree of accuracy.

Earlier this year, the mouse consortium announced that it had assembled the draft sequence of the mouse and deposited it into public databases. The consortium's paper this week reports the initial description and analysis of this text and the first global look at the similarities and the differences in the genomic landscapes of the human and mouse. The analysis was led by the Mouse Genome Analysis Group. Below are some of the highlights.

- **Human Sequence: It's Bigger, But Is It Better?** The mouse genome is 2.5 billion DNA letters long, about 14 percent shorter than the human genome, which is 2.9 billion letters long. But bigger doesn't always mean better, say scientists. The human genome is bigger because it is filled with more repeat sequences than the mouse genome. Repeat sequences are short stretches of DNA that have been hopping around the genome by copying and inserting themselves into new regions. They are not thought to have functional significance. The mouse genome, it seems, is more fastidious with its housecleaning than the human. Although it is actually accumulating repeat sequence at a greater rate than humans, it is losing them at an even greater rate.
- **Shuffling the Chapters of an Ancestral Book.** The mouse and human genomes descended from a common ancestor some 75 million years ago. Since then there has been considerable shuffling of the DNA order both within and between chromosomes. Nonetheless, when scientists compared the human and mouse genomes, they discovered that more than 90 percent of the mouse genome could be lined up with a region on the human genome. That is because the gene order in the two genomes

is often preserved over large stretches, called conserved synteny. In fact, the mouse genome could be parsed into some 350 segments, or chapters for which there is a corresponding chapter in the human genome. For example, chromosome 3 of the mouse has chapters from human chromosomes 1, 3, 4, 8 and 13, and chromosome 16 of the mouse has chapters from human chromosome 3, 21, 22 and 16.

- **Heavy Editing at the Level of Sentences.** Although virtually all of the human and mouse sequence can be aligned at the level of large chapters, only 40 percent of the mouse and the human sequences can be lined up at the level of sentences and words. Even within this 40 percent, there has been considerable editing, as evolution relentlessly tinkers with the genome. The change is so great in most places that only with very sensitive tools can scientists discern the relationships.
- **Preserving the Gems.** Despite the heavy editing, about 5 percent of the genome contains groups of DNA letters that are conserved between human and mouse. Because these DNA sequences have been preserved by evolution over tens of millions of years, scientists infer that they are functionally important and under some evolutionary selection. Interestingly, the proportion of the genome comprised by these functionally important parts is considerably higher than what scientists had expected. In particular, it is about three times as much as can be explained by protein-coding genes alone. This implies that the genome must contain many additional features (such as untranslated regions, regulatory elements, non-protein coding genes, and chromosomal structural elements) that are under selection for biological function. Discovering their meaning will be a major goal for biomedical research in the coming years.
- **The Gene Number.** When the human genome consortium concluded last year that the human sequence contains only 30,000 to 40,000 protein-coding genes, the news elicited a collective international gasp. Humans, it seems, have only about twice as many genes as the worm or the fly, and fewer genes than rice. Many wondered how human complexity could be explained by such a paucity of genes. The prediction has since been the subject of debate with some researchers suggesting much higher gene counts. The human-mouse comparison will likely put the yearlong speculation to rest, indicating that if anything, the gene numbers may be at the low end of the range. Today's paper suggests that the mouse and the human genomes each seem to contain in the neighborhood of 30,000 protein coding genes.
- **Sex, Smell and Infectious Disease.** Although the mouse and the human contain virtually the same set of genes, it seems that some families of genes have undergone expansion - or multiplied - in the mouse lineage. These involve genes related to reproduction, immunity and olfaction, suggesting that these physiological systems have been the focus of extensive innovation in rodents. It seems that sex, smell, and pathogens are most on the mouse's evolutionary mind. Scientists do not yet know the reasons for this, but they speculate that a shorter generation time, changes in living environment, lack of verbal and visual cues, and differences in reproduction may account for this.
- **Uneven Landscape of the Genomes.** Since the two species diverged, the ancestral text has changed considerably, with substitutions occurring in both species. Twice as many of these substitutions have occurred in the mouse compared with the human lineage. A great surprise is that mutation rates seem to vary across the genome in ways that cannot be explained by any of the usual features of DNA.
- **Empowering Mouse as a Disease Model.** The laboratory mouse has long been used to study human diseases. There are more than a hundred mouse models of Mendelian disorders, where a mutation in mouse counterparts of human disease genes results in a constellation of symptoms highly reminiscent of the human disorder. But there are many more such models to be found, and the availability of the mouse genome sequence will make their discovery only a few "mouse" clicks away. Furthermore, hundreds of additional mouse models of non-Mendelian diseases such as epilepsy, asthma, obesity, colon cancer, hypertension, and diabetes, which have been more difficult to pin down, will now be much more accessible to the tools of the molecular geneticist.

- **Understanding the Mouse.** The mouse genome sequence will also open new paths of scientific endeavor aimed at understanding how the mouse genome directs the biology of this mammal. Scientists will no longer be working on genes in isolation, but will view individual genes in the context of all other related genes and in the context of a whole organism. They will be able to study many, even all, genes simultaneously, speeding the understanding of the mouse in molecular terms. Scientists say such molecular understanding of the mouse will be essential to realize the full benefits of the human genome sequence.

The sequence information from the mouse consortium has been immediately and freely released to the world, without 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 key information services to biotechnologists.

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

"The mouse genome is a great resource for basic and applied medical research, meaning that much of what was done in a lab can now be done through the Web. Researchers can access this information through www.ensembl.org, where all the information is provided with no restriction," says Ewan Birney, Ph.D., Ensembl coordinator at the European Bioinformatics Institute.

Technology, Assembly and Analysis

The mouse sequence reported in today's journal is a high-quality draft sequence. The consortium plans to produce a "finished" version with the remaining gaps filled in and errors resolved. This phase will proceed using clone-based, or hierarchical, sequencing using the publicly available mouse genome clone map. The quality of the working draft sequence far exceeds the consortium's original expectations for this stage and was completed much sooner than initially expected, reflecting the tremendous efficiencies gained in sequencing and computational technologies in the past few years. Methods for efficient sequencing of large genomes continue to advance dramatically, and it is now becoming possible to sequence genomes much more rapidly.

The mouse sequencing strategy combines the features of the clone-based-hierarchical-shotgun and whole-genome-shotgun strategies. In the hierarchical-shotgun approach, individual large DNA fragments of known position are subjected to shotgun sequencing (shredded into small fragments that are sequenced and then reassembled in the basis of sequence overlaps). In the whole-genome-shotgun approach, the entire genome is shredded into small fragments that are sequenced and put back together on the basis of sequence overlaps. The scientists used data from more than 33 million individual sequencing experiments. Using two different computer systems, called genome assemblers, the team reconstructed the 33 million individual fragments into a draft sequence. These whole-genome assemblers, called ARACHNE and PHUSION were developed at the Whitehead Institute and at the Sanger Institute, respectively.

These long stretches of sequence, called contigs, were then linked into larger fragments called supercontigs of a typical length of 16.9 million base pairs. These supercontigs were then anchored to the mouse genetic and BAC clone maps. Finally, adjacent supercontigs were joined into even larger ultracontigs on the basis of other linking information. In the end, nearly the entire chromosomal sequence is contained in a mere 88 ultracontigs with a typical size of 50 megabases each.

Once assembled, the Ensembl database team, based at the EBI and Sanger Institute and funded primarily by the Wellcome Trust, coordinated the annotation of the assembled genome with interesting features, including genes. Ensembl automatically estimates where the genes are, primarily by finding similarities with known genes - both in mouse and in other organisms, such as human.

The work was then taken up by the mouse genome analysis group, comprised of genome analysis experts from 27 institutions in six countries. This "virtual genome analysis center" convened over the Internet and by regular conference calls to interpret and analyze the mouse-human comparison.

"It's been very exciting to see how much we can learn from comparing the mouse genome to the human genome and yet we see how much more there is to study," says Kerstin Lindblad-toh, Ph.D., lead author of the Nature paper and Senior Program Manager Mouse Genome at Whitehead Institute.

What's Next?

The consortium's ultimate goal is to produce a completely "finished" mouse sequence - with no gaps and 99.99 percent accuracy. Researchers expect this will take two or three years. Although the near-finished version is adequate for most biomedical research, the Human Genome Project (HGP) has made a commitment to filling all gaps and resolving all ambiguities, as it did with the human sequence. The human sequence is expected to be finished by April 2003.

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

Next in line for sequencing are the chimpanzee, the chicken, the cow, the dog, several species of fungi, a sea urchin, the honey bee and two simple organisms commonly used in laboratory studies (*Oxytricha trifallax* and *Tetrahymena thermophila*). Each organism will shed unique insights into human health and disease.

Comparative genomics will also offer scientists insights into important regions in the sequence that perform regulatory functions, such as switching on or off a gene or controlling its expression.

Finally, the sequence will serve as a foundation for a broad range of functional genomic tools to help biologists to probe the function of the genes in a more systematic manner. Development of such genomic tools will be one of the major thrusts for biologists in the next decade, according to the scientists.

All of the results from this analysis can be found at several Web sites, including at the [European Bioinformatics Institute](http://www.ebi.ac.uk) [mouse.ensembl.org], at the [National Center for Biotechnology Information](http://www.ncbi.nlm.nih.gov) at the National Library of Medicine, and at the [University of California, Santa Cruz](http://genome.ucsc.edu) [genome.ucsc.edu].

The \$130 million effort has been funded by government agencies and public charities in the various countries, including the NIH's NHGRI, the National Cancer Institute, the National Institute on Deafness and Other Communication Disorders, the National Institute of Diabetes and Digestive and Kidney Disease, the National Institute of Neurological Disorders and Stroke, and the National Institute of Mental Health; the U.S. Department of Energy; The Wellcome Trust and the Medical Research Council in England; and agencies in Canada, Japan, Germany, Switzerland and Spain. In addition, funding for the draft sequence was also provided by GlaxoSmithKline, The Merck Genome Research Institute and Affymetrix, Inc.

NHGRI is one of the 27 institutes and centers at the NIH, which is an agency of the Department of Health and Human Services (DHHS). The NHGRI Division of Extramural Research supports grants for research and for training and career development at sites nationwide. Additional information about NHGRI can be found at its Web site, www.genome.gov.

Contact:

Geoff Spencer
NHGRI
Phone: (301) 402-0911

Seema Kumar
Whitehead Institute
Phone: (617) 252-1420

[⬆ Top of page](#)

Last Updated: March 2006