PublisherInfo		
PublisherName	:	BioMed Central
PublisherLocation	:	London
PublisherImprintName	:	BioMed Central

Human genome database unveiled

ArticleInfo		
ArticleID	:	4939
ArticleDOI	:	10.1186/gb-spotlight-20040421-01
ArticleCitationID	:	spotlight-20040421-01
ArticleSequenceNumber	:	291
ArticleCategory	:	Research news
ArticleFirstPage	:	1
ArticleLastPage	:	3
ArticleHistory	:	RegistrationDate : 2004–4–21 OnlineDate : 2004–4–21
ArticleCopyright	:	BioMed Central Ltd2004
ArticleGrants	:	
ArticleContext	:	130594411

Tabitha M Powledge
Email: tam@nasw.org

A broad-ranging international collaboration reports in PLoS Biology the validation of more than 21,000 human gene candidates, including more than 5,100 new genes and nearly 300 noncoding RNA genes. The researchers also identified hundreds of polymorphic microsatellite repeats and many thousands of single nucleotide polymorphisms that could alter proteins and cause disease.

The Japan-based collaboration also reveals a novel tool for making sense of the human genome. The H-Invitational Database (H-InvDB), based on more than 41,000 full-length cDNAs, comprises the most elaborate annotation of human genes to date. It includes descriptions of gene structures, novel alternative splicing isoforms, functional domains, subcellular localizations, metabolic pathways, and comparisons with mouse cDNA. The database is free and open to all.

"This is a pretty important contribution," Sean Eddy, who heads the Howard Hughes Medical Institute genetics lab at Washington University in St. Louis, told us. Full-length cDNA sequences such as those in H-InvDB have always been the gold standard for annotating genes, he noted. "The RIKEN group and the FANTOM consortium have already provided such a resource for mouse, and the model organism folks (fly especially) are working on their own full-length cDNA resources. This paper is the long-awaited contribution for human full-length cDNA sequencing," said Eddy, who was not involved in the study.

Takashi Gojobori of the Japan Biological Information Research Center told us that, among its other advantages, H-InvDB will help increase confidence that genes identified by computer prediction can be validated experimentally. Researchers can request full-length cDNA clones through the database, said Gojobori - also at the DNA Data Bank of Japan at the National Institute of Genetics in Shizuoka, which heads the project - and use them to determine protein function.

In addition to the cDNA annotations, H-InvDB contains two other subdivisions: H-Angel, a database of gene expression patterns for more than 19,000 loci, and DiseaseInfo Viewer, a database of known disease-related genes and loci.

"This landmark collection of full-length human cDNA sequences will give us new insights into the human genome in several ways," said Terry Gaasterland, associate professor at Rockefeller University. She noted "striking" differences between H-InvDB and the reference human sequence (RefSeq) at the US National Center for Biotechnology Information. "Over 5,000 genes present in this collection are not represented in RefSeq, while nearly 3,500 curated RefSeq mRNAs are not captured in this collection. This underscores the importance of generating multiple, large cDNA datasets from many tissue and developmental stage libraries. Each new dataset like this helps to refine the assembly of the human genome." Gaasterland, who was not involved in the study, pointed out that the new paper suggests that up to 4% of the sequences in RefSeq may be missing or incorrect.

"It's a very good thing when people revisit genomes and make the information more complete and more accurate," said Eric Jakobsson, director of the Computational Biology Center at the US National Institute of General Medical Sciences and who was also not involved in the study. "It will reduce forever the noise in the inferences that we draw about them."

The project involved 152 scientists at 40 institutions in Australia, Brazil, China, France, Germany, Japan, South Africa, South Korea, Sweden, Switzerland, the United Kingdom, and the United States. Consortium members met several times and held two large annotation jamborees in Tokyo in 2002 and 2003.

"This dataset represents an enormous annotation and analysis effort. It opens the door to new levels of human genome analysis, including a more comprehensive comparison with mouse and rat, especially in the non-protein-coding regions, as well as deeper insights into the prevalence of alternatively spliced forms of genes," said Gaasterland.

References

1. *PLoS Biology*, [http://www.plosbiology.org/]

2. H-Invitational Database, [http://www.h-invitational.jp]

3. Sean Eddy, [http://www.genetics.wustl.edu/molgen/eddy.html]

4. Takashi Gojobori, [http://www.ddbj.nig.ac.jp/]

5. Terry Gaasterland, [http://genomes.rockefeller.edu/]

6. RefSeq, [http://www.ncbi.nlm.nih.gov/RefSeq/]

7. Eric Jakobsson, [http://www.nigms.nih.gov/about_nigms/cbcb.html]

This PDF file was created after publication.