The H-Invitational Database (H-InvDB), a comprehensive annotation resource for human genes and transcripts*

Received September 16, 2007; Revised October 20, 2007; Accepted October 22, 2007

ABSTRACT
Here we report the new features and improvements in our latest release of the H-Invitational Database (H-InvDB; http://www.h-invitational.jp/), a comprehensive annotation resource for human genes and transcripts. H-InvDB, originally developed as an integrated database of the human transcriptome based on extensive annotation of large sets of full-length cDNA (FLcDNA) clones, now provides annotation for 120,558 human mRNAs extracted from the International Nucleotide Sequence Databases (INSD), in addition to 54,978 human FLcDNAs, in the latest release H-InvDB_4.6. We mapped those human transcripts onto the human genome sequences (NCBI build 36.1) and determined 34,699 human gene clusters, which could define 34,057 (98.1%) protein-coding and 642 (1.9%) non-protein-coding loci; 858 (2.5%) transcribed loci overlapped with predicted pseudogenes. For all these transcripts and genes, we provide comprehensive annotation including gene structures, gene functions, alternative splicing variants, functional non-protein-coding RNAs, functional domains, predicted sub cellular localizations, metabolic pathways, predictions of protein 3D structure, mapping of SNPs and microsatellite repeat motifs, co-localization with orphan diseases, gene expression profiles, orthologous genes, protein–protein interactions (PPI) and annotation for gene families. The current H-InvDB annotation resources consist of two main views: Transcript view and Locus view and eight sub-databases: the DiseaseInfo Viewer, H-ANGEL, the Clustering Viewer, G-integra, the TOPO Viewer, Evola, the PPI view and the Gene family/group.

INTRODUCTION
Human transcripts represent a biologically and functionally rich format for examining the structure of human genes and alternative splicing isoforms. In particular, cloning and sequencing of full-length cDNAs (FLcDNAs) that cover all exons but no introns can facilitate the precise determination of human gene structure (1). Studies on human transcripts have thus been systematically and extensively carried out to draw the outline of the human transcriptome (2–6). The human transcriptome consists of protein-coding mRNAs and non-coding functional RNAs. Analysis of these sequences will provide insights into how genomic information is transformed into higher order biological phenomena. By comparative analysis of the transcriptome with the human genome, we will be able to determine the transcribed regions of the genome and better understand the regulatory machinery of transcription (7, 8). It is therefore of great significance to collect information about human transcripts as well as their annotations. We thus held the first international workshop entitled ‘Human Full-length cDNA Annotation Invitational’ (abbreviated as H-Invitational or H-Inv) in Tokyo, Japan from 25th August to 3rd September 2002, and constructed a novel, integrative database of the human transcriptome, called H-InvDB (9,10). This consists of the annotation of 42,421 human FLcDNAs, collected from six high-throughput producers of human FLcDNAs in the world human gene collections.

To cover the increased number of human FLcDNAs since the initial release of H-InvDB, we held the second international annotation meeting entitled ‘H-Invitational 2 Functional Annotation Jamboree’ (abbreviated as H-Invitational 2 or H-Inv2) in Tokyo, Japan from 15th to 20th November 2003. The second major release of H-InvDB (release 2.0) was based on the annotation carried out at the H-Inv2 annotation jamboree. After H-Inv2, we initiated the Genome Information Integration Project (GIIP) and held the third and fourth annotation meetings in October 2005 and October 2006. The products of those two annotation meetings comprised releases 3.0 and 4.0 of H-InvDB. The increases in the number of entries in H-InvDB are summarized in Table 1.

THE ANNOTATION IN OUR LATEST UPDATE, H-InvDB 2007
In our latest release H-InvDB_4.6, we annotated 120,558 human mRNAs extracted from the International Nucleotide Sequence Databases (INSD) in addition to 54,978 human FLcDNAs that were available on 15th June 2006. We mapped those human transcripts onto the human genome sequences (NCBI build 36.1) and determined 34,699 human gene clusters, which could define 34,057...
protein-coding and 643 (1.9%) non-protein-coding loci, while 858 (2.5%) transcribed loci overlapped with predicted pseudogenes. We basically followed the mapping technique we described previously (9,10). We updated annotation for the mitochondrial transcripts since the previous major release, H-InvDB_4.0, which resulted in a slightly decreased number for the transcripts and clusters. Then we assigned a standardized functional annotation to each H-Inv transcript by human curation, based on the results of similarity searches and InterProScan (11). The numbers of manually curated human proteins in each category are summarized in Table 2.

For these transcripts and genes, we provide comprehensive annotation including descriptions of their gene structures, alternative splicing isoforms, functional non-protein-coding RNAs, functional domains of proteins, predicted sub cellular localizations, metabolic pathways, predictions of protein 3D structure, mapping of SNPs and microsatellite repeat motifs, co-localization with orphan diseases, gene-expression profiles, orthologous genes and evolutionary features in model animals, protein–protein interaction (PPI) and annotation for gene families. We have also annotated several new features related to transcript quality.

**NEW ANNOTATED FEATURES IN H-InvDB**

**Classification of ncRNA**

We annotated the transcripts that do not have homology to known protein-coding genes or InterPro-domain-containing genes as non-protein-coding transcript candidates. We classified 1216 non-protein-coding transcripts into ‘Identical to known ncRNA’ (124), ‘Similar to known ncRNA’ (74) and ‘Putative ncRNA’ (1018) by homology with known ncRNA databases and discrimination analysis.

**Sequence quality features: nonsense-mediated decay (NMD), read-through, reverse orientation**

A total of 269 transcripts were annotated as candidates of read-through and 2731 as targets of NMD by the extended sequence quality annotation.

**Category VII: pseudogene candidates**

To annotate transcribed pseudogene candidates, we did the following: First, we filtered out the functional protein-coding genes by only targeting representative category II transcripts and those identified to have frame shifts and/or nonsense mutations; Second, we predicted transcribed pseudogene candidates based on a support vector machine (SVM) method. In the current release, we annotated 1112 transcribed pseudogene candidates (Category VII).

**Annotation of gene families/groups**

We annotated four selected gene families/groups: T-cell receptor (TCR), Immunoglobulin (Ig), Major Histocompatibility Complex (MHC) or Human Leukocyte Antigen (HLA) and Olfactory receptor (OR) using the original pipeline based on sequence analysis against genome and protein databases complemented by a text-mining approach. In the current release, we identified 15 TCR, 21 Ig, 72 MHC and 122 OR gene clusters.

All the annotation items and features of H-Inv transcript sequences are stored and shown in the main views or sub-databases in H-InvDB.

**COMPREHENSIVE ANNOTATION RESOURCES IN H-InvDB**

The current H-InvDB annotation resources consist of two main views, Transcript view and Locus view, and eight sub-databases: the DiseaseInfo Viewer, H-ANGEL, the Clustering Viewer, G-integra, the TOPO Viewer, Evola, the PPI view and the Gene family/group view with the appropriate cross-links. An overview of the comprehensive annotation resources of the human gene and transcripts in H-InvDB is shown in Figure 1.
Transcript view

The transcript view shows all the annotation of the H-Inv transcript in 12 section tabs: (i) gene structure, (ii) gene function, (iii) gene ontology, (iv) predicted CDS, (v) functional motif, (vi) sub cellular localization, (vii) protein structure information, (viii) gene expression, (ix) disease/pathology, (x) evolutionary information, (xi) polymorphism (SNP, indel and microsatellite) and...
interspersed repeat information and (xii) transcript and sequence quality information. As seen in the example of a transcript view shown in Figure 1, this view also has links to many external public databases including DDBJ/EMBL/GenBank, RefSeq, UniProtKB, HGNC, InterPro, Ensembl, EntrezGene, PubMed, dbSNP, GO and GTOP and to web sites of the original data producers of the FLcDNA clones and sequences including the Chinese National Human Genome Center (CHGC), German cDNA Consortium (DKFZ/MIPS), Helix Research Institute, Inc. (HRI), the Institute of Medical Science in the University of Tokyo (IMSUT), the Kazusa DNA Research Institute (KDIR), the Mammalian Gene Collection (MGC/NCI) and NEDO. This view was previously known as the cDNA view (mRNA view).

**Locus view**

The Locus view shows all the(annotation of a locus in six section tabs: (i) gene structure and location in the human genome, (ii) gene function, (iii) alternative splicing pattern, (iv) gene expression, (v) disease/pathology and (vi) cluster member information. As seen in the example of a Locus view shown in Figure 1, it shows links to external public databases including DDBJ/EMBL/GenBank, RefSeq, EntrezGene, GeneCards, HGNC and OMIM.

**DiseaseInfo Viewer**

The DiseaseInfo Viewer is a database of known and orphan genetic diseases and their relation to H-Inv clusters with EntrezGene and OMIM cross-links. The DiseaseInfo Viewer provides two kinds of disease information related to H-Inv clusters: known disease-related genes and co-localized orphan diseases. An orphan disease is defined as a disease mapped on a chromosomal region, but for which the responsible gene has not been identified yet. Co-localization does not necessarily mean a direct relationship between gene and disease; however, genes that are cytogenetically co-localized with a disease could be possible candidate genes for that disease. The co-localized H-Inv clusters are chosen by computing the physical range of each cytogenetic band with a 1 Mbp margin.

**Human anatomic gene expression library (H-ANGEL)**

H-ANGEL is a database of expression patterns that we constructed to obtain a broad outline of such patterns for human genes (12). We collected gene-expression data in normal and adult human tissues that were generated by three types of methods and in seven different platforms, including: iAFLP, a PCR-based quantitative expression profiling method; DNA arrays (long oligomers, short oligomers and cDNA microarrays); and cDNA sequence tags (SAGE, EST, BodyMap and MPSS). The H-ANGEL database comprises the largest and most comprehensive collection of gene expression patterns so far, which also provides a classification of human genes in terms of their expression.

**Clustering Viewer**

The Clustering Viewer facilitates the comparisons of different clustering. It allows users to see whether H-Inv transcripts are consistently clustered by different clustering methods. It also displays multiple alignments of transcripts by using CLUSTALW (13). The Clustering Viewer shows all the member transcripts of an H-Inv cluster to which a query sequence belongs.

**G-integra**

G-integra is an integrated genome browser, in which we can examine the genomic structures of the transcripts. As seen in an example view in Figure 1, the location in the human genome and gene structure of H-Inv transcript (green), and the corresponding RefSeq and Ensembl entries are shown. The structures of the genes and transcripts for 11 non-human species, *Pan troglodytes* (chimpanzee), *Macaca* sp. (macaque), *Mus musculus* (mouse), *Rattus norvegicus* (rat), *Canis familiaris* (dog), *Bos taurus* (cow), *Monodelphis domestica* (opossum), *Gallus gallus* (chicken), *Danio rerio* (zebrafish), *Tetraodon nigroviridis* (tetraodon) and *Takifugu rubripes* (fugu) can be optionally displayed for comparison. Other options allow the, the results of gene prediction programs such as GenScan (14), HMMgene (15), FGENESh (16) and JIGSAW (17) to be displayed.

**TOPO Viewer**

The TOPO Viewer is a tool for viewing subcellular targeting signals predicted by TargetP (18) and the presence of transmembrane helices predicted by SOSUI (19) and TMHMM (20). The probabilities that a protein may be delivered to up to nine distinct sub cellular locations are predicted by WoLF PSORT (21). TargetP predicts whether a protein contains a signal peptide, a mitochondrial targeting signal or any other type of signal. The TOPO Viewer consists of four tab pages: TABLE, MAP, FILE and GFP. The TABLE tab page displays the prediction results for all the programs used.

**Evola**


**PPI view**

The PPI view displays H-invDB human PPI information at http://www.jbirc.aist.go.jp/hinv/ppi/. We collected PPI data from five databases; BIND, DIP, MINT, HPRD and IntAct, removed redundancies of the PPI data among the
databases based on their sequence similarities and integrated them with the H-Invitational proteins.

**Acknowledgements**

We acknowledge all the members of the H-Invitational 2 consortium and Genome Information Integration Project (GIIP), especially the staffs of JBIC for construction of H-InvDB, Ryo Aono, Tomohiro Endo, Yukie Makita, Hiromi Kubooka, Yuji Shinso, Harutoshi Maekawa, Yasuhiro Fukunaga, Hajime Nakaoka, Yoshito Ueki, Yoshihide Mimura, Ryuzou Matsumoto, Seigo Hosoda, Yo Takahashi, Taichiroy Sugisaki, Hiroki Hokari, Hiroaki Kawashima, Yasuhiro Imamizu, Makoto Ogawa for their technical assistance. This research is financially supported by the Ministry of Economy, Trade and Industry of Japan (METI), the Ministry of Education, Culture, Sports, Science and Technology of Japan (MEXT) and the Japan Biological Informatics Consortium (JBIC). Also, this work is partly supported by the Research Grant for the RIKEN Genome Exploration Research Project from MEXT to Y.H. and the Grant for the RIKEN Frontier Research System, Functional RNA research program. Funding to pay the Open Access publication charges for this article was provided by JBIC.

**Conflict of interest statement.** None declared.

**References**

LIST OF AUTHORS FOR THE GENOME INFORMATION INTEGRATION PROJECT AND H-INVITATIONAL 2 CONSORTIUM


1Japan Biological Information Research Center, Japan Biological Informatics Consortium, 2Biological Information Research Center, National Institute of Advanced Industrial Science and Technology, Tokyo, 3Graduate School Medicine, Dentistry and Pharmaceutical Sciences, Okayama University, Okayama, 4DNA Chip Research Inc., Kanagawa, 5Hokkaido University, Hokkaido, Japan, 6Centre for Comparative Genomics, Murdoch University, WA, Australia, 7Biotechnology Research Center, The University of Tokyo, 8Hitachi Software Engineering Co., Ltd., 9Mitsubishi Kagaku Institute of Life Sciences, 10Fujitsu Limited, Tokyo, 11Hitachi, Co., Ltd., Saitama, 12Japan Science and Technology Agency, 13NEC Soft, Ltd., 14Mitsui Knowledge Industry Co., Ltd., Tokyo, 15National Institute of Agrobiological Sciences, Ibaraki, 16BITS Co., Ltd., Shizuoka, 17Tokyo Institute of Psychiatry, Tokyo, 18DNA-NACOM Co., Ltd., Chiba, 19C's Lab Co., Ltd., Hokkaido, 20Center for Information Biology and DNA Data Bank of Japan, National Institute of Genetics, Shizuoka, 21Tokyo University of Science, Chiba, Japan, 22University of Dublin, Trinity College, Dublin, Ireland, 23Mitsubishi Space Software Co., Ltd., Ibaraki, 24Division of Population Genetics, National Institute of Genetics, Shizuoka, Japan, 25EMBL Outstation-Hinxton, European Bioinformatics Institute, Cambridge, UK, 26Bioinformatics Research Center, Medical College of Wisconsin, WI, USA, 27Center for Computational Science and Engineering, Japan Atomic Energy Agency, Kyoto, 28Central Research Laboratory, Hitachi Ltd., 29Reverse Proteomics Research Institute, CO., Ltd., 30Computational Biology Research Center, National Institute of Advanced Industrial Science and Technology, Tokyo, 31Department of Human Gene, Kazusa DNA Research Institute, Chiba, Japan, 32Department of Molecular Genetics, Weizmann Institute of Science, Rehovot, Israel, 33Genexpres, Functional Genomics and Systems Biology for Health (CNRS and Pierre & Marie Curie University - Paris VI), Villejuif, France, 34Sino-French Laboratory in Life Sciences and Genomics, Shanghai, China, 35Centre de Génétique Moléculaire, CNRS and Gil/Oryx DNA Microarray Platform, Gif/Yvette, 36Laboratory of Genomes Functional Exploration, CEA, DSV, IRCCM, Evry, France, 37Genomic Sciences Center, RIKEN Yokohama Institute, Kanagawa, 38Genome Science Laboratory, Discovery and Research Institute, RIKEN Wako Institute, Saitama, Japan, 39GSF - National Research Center for Environment and Health, Institute for Bioinformatics,
Neuherberg, Germany, 40Idaho State University, ID, USA, 41Institute for Chemical Research, Kyoto University, Kyoto, Japan, 42Institute of Bioinformatics, University of Muenster, Muenster, Germany, 43Kazusa DNA Research Institute, Chiba, Japan, 44Korea Research Institute of Bioscience & Biotechnology, Taejeon, Korea, 45Ludwig Institute for Cancer Research, Sao Paulo, Brazil, 46Medical Education and Biomedical Research Facility, University of Iowa, IA, USA, 47Medical Research Institute, Tokyo Medical and Dental University, Tokyo, Japan, 48Molecular Genome Analysis, German Cancer Research Center, Heidelberg, Germany, 49Nagahama Institute of Bio-Science and Technology, Shiga, Japan, 50National Cancer Institute, National Institutes of Health, MD, 51National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, MD, USA, 52National Engineering Center for Biochips at Shanghai, Shanghai, China, 53Ochanomizu University, Tokyo, 54Photon Medical Research Center, Hamamatsu University School of Medicine, Shizuoka, 55Plant Science Center, RIKEN Yokohama Institute, Kanagawa, 56Harvard Medical School, MA, USA, 57Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai, China, 58Center for Genomics and Bioinformatics, Karolinska Institute, Stockholm, Sweden, 59Soongsil University, Seoul, Korea, 60State Key Laboratory of Medical Genomics, Shanghai Institute of Hematology, Rui Jin Hospital, Shanghai Jiao Tong University School of Medicine, 61Chinese National Human Genome Center at Shanghai, Shanghai, China, 62Swiss Institute of Bioinformatics, Geneva, Switzerland, 63The Institute of Medical Science, The University of Tokyo, Tokyo, Japan, 64The Pennsylvania State University, PA, USA, 65The South African National Bioinformatics Institute, University of Western Cape, Cape Town, South Africa, 66The Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Cambridge, UK, 67University of Cincinnati, OH, 68Children’s Memorial Research Center, Northwestern University, Feinberg School of Medicine, USA, 69University of Naples “Federico II”, Naples, Italy and 70Department of Medical Genome Sciences, Graduate School of Frontier Sciences, The University of Tokyo, Tokyo, Japan

To whom correspondence should be addressed. Tel: +81-3-3599-8800; Fax: +81-3-3599-8801; E-mail: t.imanishi@aist.go.jp
Correspondence may also be addressed to Takashi Gojobori. Tel: +81-55-981-6847; Fax: +81-55-981-6848; Email: tgojobor@genes.nig.ac.jp