Published online 18 December 2007 *Nucleic Acids Research, 2008, Vol. 36, Database issue* ***D793–D799***
doi:10.1093/nar/gkm999

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

*A complete list of authors appears at the end of this article.

Table 1. Statistics of H-InvDB entries

H-InvDB release	Date of release	Number of transcripts (HIT)	Number of gene clusters (HIX)	Number of proteins (HIP)	Human genome	Date of sequence data-fix
1.0	2004/4/20	41 118	21 037	–	NCBI build 34.1	2002/7/15
2.0	2005/8/31	56 419	25 585	–	NCBI build 34.1	2003/9/1
3.0	2006/3/31	167 992	35 005	–	NCBI build 35.1	2005/3/1
4.0	2007/3/30	175 542	34 701	116 228	NCBI build 36.1	2006/6/15
4.6	2007/9/27	175 536	34 699	116 142	NCBI build 36.1	2006/6/15

Table 2. Statistics of manually curated representative H-Inv proteins

Category	Definition	Number of representative HITs	%
I	Identical to known[a] human protein (≥98% identity, =100% coverage)	12 404	36.42
II	Similar to known[a] protein (≥50% identity, ≥50% coverage)	3165	9.29
III	InterPro domain containing protein	3056	8.97
IV	Conserved hypothetical protein	4210	12.33
V	Hypothetical protein	5124	15.05
VI	Hypothetical short protein (20–79 amino acids)	5250	15.42
VII	Pseudogene candidates	858	2.52
Total		34 057	100

[a]'Known' proteins are experimentally validated proteins in literatures.

(98.1%) 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.

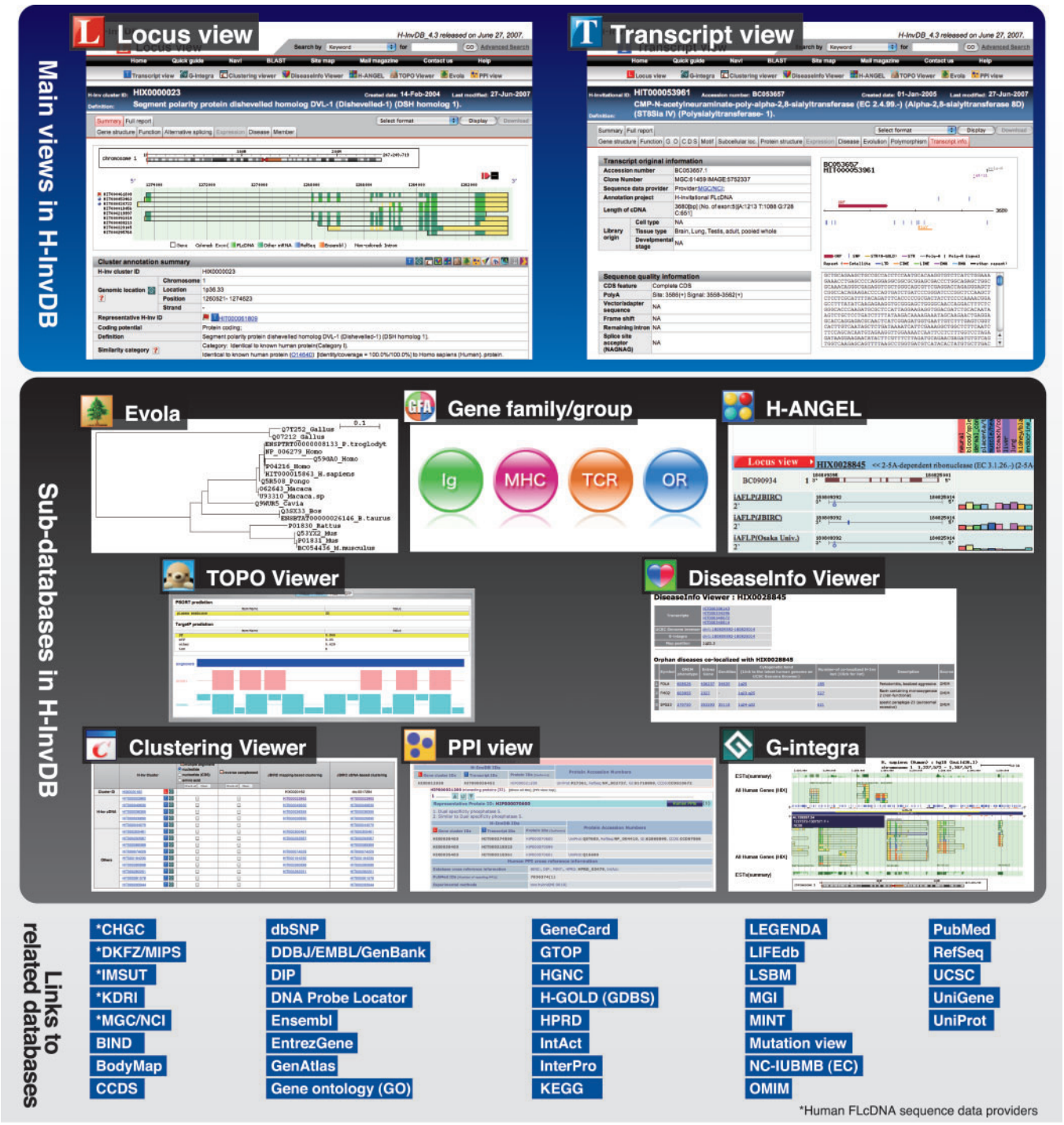

Figure 1. H-InvDB: overview of the comprehensive annotation resource for the human genes and transcripts. The current H-InvDB annotation resources consist of two main views, Transcript view and Locus view, and eight sub-databases: the DiseaseInfo Viewer, H-ANGEL, the Clustering Viewer, G-integra, the TOPO Viewer, Evola, the PPI view and the Gene family/group view. The Transcript view and the Locus view are the main viewers to display the annotation of each H-Invitational transcript (HIT) and H-Invitational cluster (HIX). The DiseaseInfo Viewer, H-ANGEL, the Clustering Viewer, G-integra, the TOPO Viewer, Evola, the PPI view and the Gene family/group view are sub-databases to provide detailed annotation for each annotation feature. The links to related databases are provided from the appropriate viewers.

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 (KDRI), 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

Evola is a database of evolutionary annotation of human genes (22). It provides sequence alignments and phylogenetic trees of manually curated orthologous genes among human and 11 model organisms, *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* (zebra fish), *Tetraodon nigroviridis* (tetraodon) and *Takifugu rubripes* (fugu). Sequence alignments and phylogenetic trees of the orthologous genes and homologous genes are shown in 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.

Gene family/Group view

The Gene family/Group view provides human-curated annotation datasets for the selected gene families/groups at http://www.jbirc.aist.go.jp/hinv/ahg-db/geneFamilyIndex.jsp. For H-InvDB release 4.0, we provided detailed annotations for four selected gene families/groups: TCR, Ig, MHC and OR. Each page provides the list of genes, gene names, definitions and links for the appropriate H-InvDB views.

H-InvDB New Identifier

We defined and assigned a unique identifier for each annotation unit, transcript, protein or cluster (7,8). The identifier for H-Invitational transcript is 'HIT', prefix HIT plus nine digit numbers (e.g. HIT000000001) and for H-Invitational cluster is 'HIX', prefix HIX plus seven digit numbers (e.g. HIX0000001). In order to identify the modification in sequence or annotation of an H-Inv entry, a version is assigned to each ID and always stated with the ID. Additionally, we now provide a new identifier for each H-Invitational protein, 'HIP', prefix HIP with nine digit numbers (e.g. HIP000000001).

H-InvDB Data Availability

H-InvDB is freely available for both academic and commercial use and can be accessed online at http://www.h-invitational.jp/(or hinv.jp). Annotated data can also be downloaded in FASTA sequence files, the original-format flat files or XML files at HTTP and FTP servers. The mirror database is also available at http://hinvdb.ddbj.nig.ac.jp/. Minor updates are released every three months and major updates are released once a year.

ACKNOWLEDGEMENTS

We acknowledge all the members of the H-Invitational 2 consortium and Genome Information Integration Project (GIIP), especially the staffs of JBIRC for construction of H-InvDB, Ryo Aono, Tomohiro Endo, Yukie Makita, Hiromi Kubooka, Yuji Shinso, Harutoshi Maekawa, Yasuhiro Fukunaga, Hajime Nakaoka, Yoshito Ueki, Yoshihide Mimiura, Ryuzou Matsumoto, Seigo Hosoda, Yo Takahashi, Taichirou 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

1. Ota,T. *et al.* (1997) Full-length cDNA project toward a high throughput functional analysis. *Microb. Comp. Genomics*, **2**, 204–205.
2. Yudate,H.T. *et al.* (2001) HUNT: launch of a full-length cDNA database from the helix research institute. *Nucleic Acids Res.*, **29**, 185–188.
3. Wiemann,S. *et al.* (2001) Toward a catalog of human genes and proteins: sequencing and analysis of 500 novel complete protein coding human cDNAs. *Genome Res.*, **11**, 422–435.
4. Strausberg,R.L. *et al.* (2002) Generation and initial analysis of more than 15,000 full-length human and mouse cDNA sequences. *Proc. Natl Acad. Sci. USA*, **99**, 16899–16903.
5. Kikuno,R. *et al.* (2002) HUGE: a database for human large proteins identified in the Kazusa cDNA sequencing project. *Nucleic Acids Res.*, **30**, 166–168.
6. Carninci,P. *et al.* (2005) The transcriptional landscape of the mammalian genome. *Science*, **309**, 1559–1563.
7. Frith,M.C. *et al.* (2006) Pseudo-messenger RNA: phantoms of the transcriptome. *PLoS Genet*, **2**, p. e23.
8. Gingeras,T.R. *et al.* (2007) Origin of phenotypes: genes and transcripts. *Genome Res.*, **17**, 682–690.
9. Imanishi,T. *et al.* (2004) Integrative annotation of 21,037 human genes validated by full-length cDNA clones. *PLoS Biol.*, **2**, 856–875.
10. Yamasaki,C. *et al.* (2005) Investigation of protein functions through data-mining on integrated human transcriptome database, H-Invitational database (H-InvDB). *Gene*, **364**, 99–107.
11. Mulder,N.J. *et al.* (2007) New developments in the InterPro database. *Nucleic Acids Res.*, **35**(Database issue), D224–D228.
12. Tanino,M. *et al.* (2005) The human anatomic gene expression library (H-ANGEL), the H-Inv integrative display of human gene expression across disparate technologies and platforms. *Nucleic Acids Res.*, **33**(Database Issue), D567–D572.
13. Thompson,J.D., Higgins,D.G. and Gibson,T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, **22**, 4673–4680.
14. Burge,C. and Karlin,S. (1997) Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.*, **268**, 78–94.
15. Krogh,A. (1997) Two methods for improving performance of an HMM and their application for gene finding. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, **5**, 179–186.
16. Salamov,A.A. and Solovyev,V.V. (2000) Ab initio gene finding in Drosophila genomic DNA. *Genome Res.*, **10**, 516–522.
17. Allen,J.E. and Salzberg,S.L. (2005) JIGSAW: integration of multiple sources of evidence for gene prediction. *Bioinformatics*, **21**, 3596–3603.
18. Emanuelsson,O. *et al.* (2000) Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. *J. Mol. Biol.*, **300**, 1005–1016.
19. Hirokawa,T., Boon-Chieng,S. and Mitaku,S. (1998) SOSUI: classification and secondary structure prediction system for membrane proteins. *Bioinformatics*, **14**, 378–379.
20. Krogh,A. *et al.* (2001) Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J. Mol. Biol.*, **305**, 567–580.
21. Horton,P. *et al.* (2007) WoLF PSORT: protein localization predictor. *Nucleic Acids Res.*, **35**(Web Server issue), W585–W587.
22. Matsuya,A. *et al.* (2008) Evola: ortholog database of all human genes in H-InvDB with manual curation of phylogenetic trees. *Nucleic Acids Res,* (in press).

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

Chisato Yamasaki[1,2], Katsuhiko Murakami[1,2], Yasuyuki Fujii[3], Yoshiharu Sato[1,2], Erimi Harada[1,2], Jun-ichi Takeda[1,2], Takayuki Taniya[1,2], Ryuichi Sakate[1,2], Shingo Kikugawa[1,2], Makoto Shimada[1,2], Motohiko Tanino[4], Kanako O. Koyanagi[5], Roberto A. Barrero[6], Craig Gough[1,2], Hong-Woo Chun[1,2], Takuya Habara[1], Hideki Hanaoka[7], Yosuke Hayakawa[1,8], Phillip B. Hilton[1,2], Yayoi Kaneko[9], Masako Kanno[1,2], Yoshihiro Kawahara[1,2], Toshiyuki Kawamura[10], Akihiro Matsuya[1,11], Naoki Nagata[12], Kensaku Nishikata[1,13], Akiko Ogura Noda[1,2], Shin Nurimoto[14], Naomi Saichi[1,2], Hiroaki Sakai[15], Ryoko Sanbonmatsu[1,2], Rie Shiba[1,2], Mami Suzuki[1,2], Kazuhiko Takabayashi[8], Aiko Takahashi[1,2], Takuro Tamura[16], Masayuki Tanaka[1,2], Susumu Tanaka[17], Fusano Todokoro[1,18], Kaori Yamaguchi[1], Naoyuki Yamamoto[1,19], Toshihisa Okido[20], Jun Mashima[20], Aki Hashizume[20], Lihua Jin[20], Kyung-Bum Lee[20], Yi-Chueh Lin[20], Asami Nozaki[20], Katsunaga Sakai[20], Masahito Tada[20], Satoru Miyazaki[21], Takashi Makino[22], Hajime Ohyanagi[20,23], Naoki Osato[20], Nobuhiko Tanaka[20], Yoshiyuki Suzuki[20], Kazuho Ikeo[20], Naruya Saitou[24], Hideaki Sugawara[20], Claire O'Donovan[25], Tamara Kulikova[25], Eleanor Whitfield[25], Brian Halligan[26], Mary Shimoyama[26], Simon Twigger[26], Kei Yura[27], Kouichi Kimura[28], Tomohiro Yasuda[28], Tetsuo Nishikawa[28,29], Yutaka Akiyama[30], Chie Motono[30], Yuri Mukai[30], Hideki Nagasaki[15,30], Makiko Suwa[30], Paul Horton[30], Reiko Kikuno[31], Osamu Ohara[31], Doron Lancet[32], Eric Eveno[33,34], Esther Graudens[33,34], Sandrine Imbeaud[33,34,35], Marie Anne Debily[33,34,36], Yoshihide Hayashizaki[37,38], Clara Amid[39], Michael Han[39], Andreas Osanger[39], Toshinori Endo[5], Michael A. Thomas[40], Mika Hirakawa[41], Wojciech Makalowski[42], Mitsuteru Nakao[43], Nam-Soon Kim[44], Hyang-Sook Yoo[44], Sandro J. De Souza[45], Maria de Fatima Bonaldo[46], Yoshihito Niimura[47], Vladimir Kuryshev[48], Ingo Schupp[48], Stefan Wiemann[48], Matthew Bellgard[6], Masafumi Shionyu[49], Libin Jia[50], Danielle Thierry-Mieg[51], Jean Thierry-Mieg[51], Lukas Wagner[51], Qinghua Zhang[34,52], Mitiko Go[53], Shinsei Minoshima[54], Masafumi Ohtsubo[54], Kousuke Hanada[55], Peter Tonellato[56], Takao Isogai[29], Ji Zhang[34,57], Boris Lenhard[58], Sangsoo Kim[59], Zhu Chen[34,60,61], Ursula Hinz[62], Anne Estreicher[62], Kenta Nakai[63], Izabela Makalowska[64], Winston Hide[65], Nicola Tiffin[65], Laurens Wilming[66], Ranajit Chakraborty[67], Marcelo Bento Soares[68], Maria Luisa Chiusano[69], Yutaka Suzuki[70], Charles Auffray[33,34], Yumi Yamaguchi-Kabata[2], Takeshi Itoh[2,15], Teruyoshi Hishiki[2], Satoshi Fukuchi[20], Ken Nishikawa[20], Sumio Sugano[2,70], Nobuo Nomura[2], Yoshio Tateno[20], Tadashi Imanishi[2,5,†] and Takashi Gojobori[2,20]

[1]Japan Biological Information Research Center, Japan Biological Informatics Consortium, [2]Biological Information Research Center, National Institute of Advanced Industrial Science and Technology, Tokyo, [3]Graduate School Medicine, Dentistry and Pharmaceutical Sciences, Okayama University, Okayama, [4]DNA Chip Research Inc., Kanagawa, [5]Hokkaido University, Hokkaido, Japan, [6]Centre for Comparative Genomics, Murdoch University, WA, Australia, [7]Biotechnology Research Center, The University of Tokyo, [8]Hitachi Software Engineering Co., Ltd., [9]Mitsubishi Kagaku Institute of Life Sciences, [10]Fujitsu Limited, Tokyo, [11]Hitachi, Co., Ltd., Saitama, [12]Japan Science and Technology Agency, [13]NEC Soft, Ltd., [14]Mitsui Knowledge Industry Co., Ltd, Tokyo, [15]National Institute of Agrobiological Sciences, Ibaraki, [16]BITS Co., Ltd., Shizuoka, [17]Tokyo Institute of Psychiatry, Tokyo, [18]DYNACOM Co., Ltd., Chiba, [19]C's Lab Co., Ltd., Hokkaido, [20]Center for Information Biology and DNA Data Bank of Japan, National Institute of Genetics, Shizuoka, [21]Tokyo University of Science, Chiba, Japan, [22]University of Dublin, Trinity College, Dublin, Ireland, [23]Mitsubishi Space Software Co., Ltd., Ibaraki, [24]Division of Population Genetics, National Institute of Genetics, Shizuoka, Japan, [25]EMBL Outstation-Hinxton, European Bioinformatics Institute, Cambridge, UK, [26]Bioinformatics Research Center, Medical College of Wisconsin, WI, USA, [27]Center for Computational Science and Engineering, Japan Atomic Energy Agency, Kyoto, [28]Central Research Laboratory, Hitachi Ltd., [29]Reverse Proteomics Research Institute, CO., Ltd., [30]Computational Biology Research Center, National Institute of Advanced Industrial Science and Technology, Tokyo, [31]Department of Human Gene, Kazusa DNA Research Institute, Chiba, Japan, [32]Department of Molecular Genetics, Weizmann Institute of Science, Rehovot, Israel, [33]Genexpres, Functional Genomics and Systems Biology for Health (CNRS and Pierre & Marie Curie University - Paris VI), Villejuif, France, [34]Sino-French Laboratory in Life Sciences and Genomics, Shanghai, China, [35]Centre de Génétique Moléculaire, CNRS and Gif/Orsay DNA Microarray Platform, Gifs/Yvette, [36]Laboratory of Genomes Functional Exploration, CEA, DSV, IRCM, Evry, France, [37]Genomic Sciences Center, RIKEN Yokohama Institute, Kanagawa, [38]Genome Science Laboratory, Discovery and Research Institute, RIKEN Wako Institute, Saitama, Japan, [39]GSF - National Research Center for Environment and Health, Institute for Bioinformatics,

Neuherberg, Germany, [40]Idaho State University, ID, USA, [41]Institute for Chemical Research, Kyoto University, Kyoto, Japan, [42]Institute of Bioinformatics, University of Muenster, Muenster, Germany, [43]Kazusa DNA Research Institute, Chiba, Japan, [44]Korea Research Institute of Bioscience & Biotechnology, Taejeon, Korea, [45]Ludwig Institute for Cancer Research, Sao Paulo, Brazil, [46]Medical Education and Biomedical Research Facility, University of Iowa, IA, USA, [47]Medical Research Institute, Tokyo Medical and Dental University, Tokyo, Japan, [48]Molecular Genome Analysis, German Cancer Research Center, Heidelberg, Germany, [49]Nagahama Institute of Bio-Science and Technology, Shiga, Japan, [50]National Cancer Institute, National Institutes of Health, MD, [51]National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, MD, USA, [52]National Engineering Center for Biochips at Shanghai, Shanghai, China, [53]Ochanomizu University, Tokyo, [54]Photon Medical Research Center, Hamamatsu University School of Medicine, Shizuoka, [55]Plant Science Center, RIKEN Yokohama Institute, Kanagawa, [56]Harvard Medical School, MA, USA, [57]Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai, China, [58]Center for Genomics and Bioinformatics, Karolinska Institute, Stockholm, Sweden, [59]Soongsil University, Seoul, Korea, [60]State Key Laboratory of Medical Genomics, Shanghai Institute of Hematology, Rui Jin Hospital, Shanghai Jiao Tong University School of Medicine, [61]Chinese National Human Genome Center at Shanghai, Shanghai, China, [62]Swiss Institute of Bioinformatics, Geneva, Switzerland, [63]The Institute of Medical Science, The University of Tokyo, Tokyo, Japan, [64]The Pennsylvania State University, PA, USA, [65]The South African National Bioinformatics Institute, University of Western Cape, Cape Town, South Africa, [66]The Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Cambridge, UK, [67]University of Cincinnati, OH, [68]Children's Memorial Research Center, Northwestern University, Feinberg School of Medicine, USA, [69]University of Naples ''Federico II'', Naples, Italy and [70]Department 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