D626–D632 *Nucleic Acids Research, 2010, Vol. 38, Database issue* *Published online 23 November 2009*
doi:10.1093/nar/gkp1020

H-InvDB in 2009: extended database and data mining resources for human genes and transcripts

Chisato Yamasaki[1], Katsuhiko Murakami[2], Jun-ichi Takeda[1], Yoshiharu Sato[1], Akiko Noda[1], Ryuichi Sakate[1], Takuya Habara[1], Hajime Nakaoka[2,3], Fusano Todokoro[2,4], Akihiro Matsuya[2,5], Tadashi Imanishi[1] and Takashi Gojobori[1,6,*]

[1]BIRC, AIST, [2]JBIC, [3]C's Lab Co. Ltd, [4]DYNACOM Co. Ltd, [5]Hitachi Ltd, [6]CIB-DDBJ, NIG Waterfront Bio-IT Research Building, 4-7 Aomi, Koto-ku, Tokyo 135-0064, Japan

Received September 16, 2009; Revised and Accepted October 19, 2009

ABSTRACT

We report the extended database and data mining resources newly released in the H-Invitational Database (H-InvDB; http://www.h-invitational.jp/). H-InvDB is a comprehensive annotation resource of human genes and transcripts, and consists of two main views and six sub-databases. The latest release of H-InvDB (release 6.2) provides the annotation for 219 765 human transcripts in 43 159 human gene clusters based on human full-length cDNAs and mRNAs. H-InvDB now provides several new annotation features, such as mapping of microarray probes, new gene models, relation to known ncRNAs and information from the Glycogene database. H-InvDB also provides useful data mining resources—'Navigation search', 'H-InvDB Enrichment Analysis Tool (HEAT)' and web service APIs. 'Navigation search' is an extended search system that enables complicated searches by combining 16 different search options. HEAT is a data mining tool for automatically identifying features specific to a given human gene set. HEAT searches for H-InvDB annotations that are significantly enriched in a user-defined gene set, as compared with the entire H-InvDB representative transcripts. H-InvDB now has web service APIs of SOAP and REST to allow the use of H-InvDB data in programs, providing the users extended data accessibility.

INTRODUCTION

We held the first international workshop entitled 'Human Full-length cDNA Annotation Invitational' (abbreviated as H-Invitational or H-Inv) in Tokyo, Japan, from 25 August to 3 September 2002, and constructed a novel, integrative database of human transcriptome called H-Invitational Database (H-InvDB; http://www.h-invitational.jp/) (1). H-InvDB is a comprehensive annotation resource of human genes and transcripts. On 20 April 2009, we marked the fifth anniversary of the opening of H-InvDB to the public. During this period, we released six major updates, namely H-InvDB 1.0(1), 2.0(2), 3.0, 4.0(3), 5.0 and 6.0. The latest release (release 6.2) provides annotations for 219 765 human transcripts in 43 159 human gene clusters based on human full-length cDNAs and mRNAs. The increases in the number of entries in H-InvDB are summarized in Table 1.

For these human transcripts, proteins and genes, we now provide several new annotation features, such as mapping of probes, new gene models, relation to known ncRNAs and glycogene information. H-InvDB now also provides useful data mining resources—'Navigation search', 'H-InvDB Enrichment Analysis Tool (HEAT)' and web service APIs. Here, we report on the extended database and data mining resources newly released in H-InvDB.

THE EXTENDED DATABASE OF H-InvDB RELEASE 6.2

In our latest release of H-InvDB release 6.2, we annotated 162 395 human mRNAs extracted from the International Nucleotide Sequence Databases (INSD)(4) in addition to 54 927 human FLcDNAs that were available on 9 May 2008. We mapped these human transcripts onto the human genome sequences (NCBI build 36.2) and determined 43 159 human gene clusters. For these human gene clusters, we defined 34 511 (80.0%) protein-coding and 7747 (17.9%) non-protein-coding loci, whereas

*To whom correspondence should be addressed. Tel: +81 3 3599 8800, Fax: +81 3 3599 8801; Email: tgojobor@genes.nig.ac.jp

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)	Annotation jamboree	
1.0	20 April 2004	41 118	21 037	–	H-Invitational 1[a]	August 2002
2.0	31 August 2005	56 419	25 585	–	H-Invitational 2 FA[a]	November 2003
3.0	31 March 2006	167 992	35 005	–	All human gene FA meeting 2005[b]	October 2005
4.0	28 March 2007	175 542	34 701	173 690	All human gene FA meeting 2006[b]	October 2006
5.0	26 December 2008	187 156	36 073	124 280	All human gene FA meeting 2007[b]	October 2007
6.0	18 December 2008	219 765	43 159	133 523		
6.2	30 March 2009	219 765	43 159	133 629		

[a]Meeting of H-Invitational project.
[b]Meeting hosted by Genome Information Integration Project (GIIP).

Table 2. Statistics of curated representative H-Inv proteins (H-InvDB release 6.2)

Category	Definition	Number of representative HITs	Percentage
I	Identical to known[a] human protein (≥98% identity, =100% coverage)	13 314	37.71
II	Similar to known[a] protein (≥50% identity, ≥50% coverage)	3380	9.57
III	InterPro domain containing protein	2584	7.32
IV	Conserved hypothetical protein	4584	12.98
V	Hypothetical protein	5203	14.74
VI	Hypothetical short protein (20–79 amino acids)	5446	15.43
VII	Pseudogene candidates	901	2.55
Total		35 303	100.00

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

901 (2.1%) transcribed loci overlapped with predicted pseudogenes. We then followed functional and further comprehensive annotation procedures as described previously (1–3). The statistics of manually curated representative human proteins are summarized in Table 2.

In H-InvDB, we now include annotation for two kinds of high-quality predicted transcripts: eHITs and pHITs. The eHIT transcripts are computationally and manually annotated gene models whose exon–intron structures are synthetically predicted by integrating the information of EST and mRNA sequences. pHIT transcripts are the novel gene candidates predicted from human genome sequences using CAGE tags and several gene prediction programs summarized using JIGSAW (5). In H-InvDB release 6.2, we provided 612 eHIT and 1831 pHIT predicted transcripts. For eHIT gene models, we assigned HIT ID prefixed 'e' (e.g. eHIT000000001) and for pHIT gene models, we assigned HIT ID prefixed 'p' (e.g. pHIT000000001). For example, pHIT000015735 is mapped on chromosome 9p13.3 and consists of 18 exons. The functional description for pHIT000015735 is 'Interleukin-11 receptor alpha chain precursor (IL-11R-alpha) (IL-11RA), Isoform HCR2' which is classified as H-InvDB similarity category I, Identical to known human protein. For pHIT000015735, HIX0153289 is assigned as cluster ID and HIP000180408 is assigned as protein ID. It is a newly identified isoform of a known UniProtKB/Swiss-Prot entry, Q14626-2, which is a soluble form of Interleukin-11 receptor alpha chain (sIL11RA). In HIX0153289, pHIT000015735 is an only member and no other human mRNA, RefSeq nor Ensembl transcripts are included, suggesting that this is a novel human transcript candidate with a support of UniProtKB/Swiss-Prot entry. An example screen shot of G-integra for pHIT000015735 is shown in Figure 1.

The H-InvDB annotation resources consist of two main views: Transcript view and Locus view, and six sub-databases: the DiseaseInfo Viewer H-ANGEL (6), G-integra, Evola (7), the PPI view and the Gene family/group view with appropriate crosslinks. Here, we describe the viewers that we have extended since our previous report (3). The new annotation features in H-InvDB are summarized in Table 3.

New features in Transcript view and Locus view

Transcript view shows all annotations of the H-Inv transcript in 12 section tabs, and Locus view shows all annotations of a locus in 6 section tabs. At the 'expression' tab in Transcript and Locus view, the mappings of microarray probes to H-InvDB data are now available. The probes of DNA Chip Research AceGene, Affymetrix GeneChip and Agilent in DNAProbe Locator (http://h-invitational.jp/DNAProbeLocator/) were mapped, related to H-InvDB entries (both to HIT and HIX), and are shown. To qualify the transcript quality, we now provide two new features, truncation (8) and Kozak consensus sequence (9) at the 'Transcript Info' tab in Transcript view. We have also integrated the annotated information of the GlycoGene Database (10) and the Functional RNA Database (11) at the 'function' tab in Transcript view using web services.

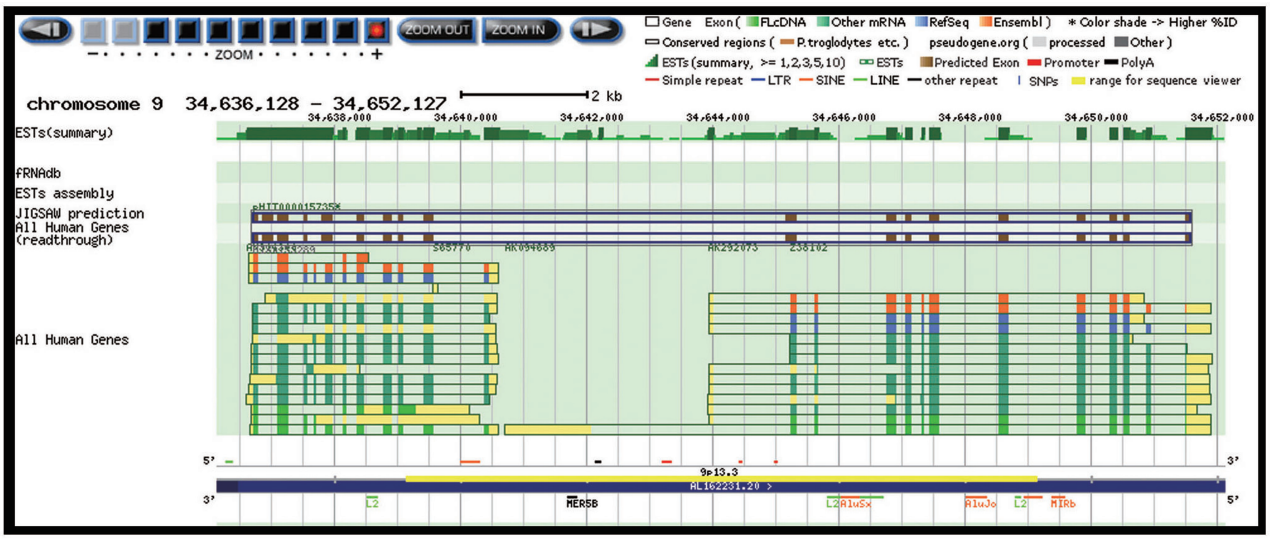

Figure 1. pHIT gene model in G-integra genome browser. An image of G-integra genome browser for a pHIT gene model, pHIT000015735, is shown (http://www.h-invitational.jp/hinv/g-integra/cgi-bin/f_genemap.cgi?id=pHIT000015735). Gene structure of pHIT000015735 is indicated by blue solid square at all human gene and JIGSAW track.

Table 3. New annotated features in H-InvDB

No.	Annotation item	Area	Available at
1	Mappings of microarray probes to H-InvDB data	Expression	'Expression' tab in Transcript view
2	New ID for gene families/groups (HIF)	Gene family	'Function' tab in Transcript view, Locus view, and Gene Family/ groups view.
3	pHIT gene models	Gene model	Transcript view, Locus view, G-integra and all the related viewers
4	eHIT gene models	Gene model	Transcript view, Locus view, G-integra and all the related viewers
5	Truncation judgment	Quality control	'Transcript Information' tab in Transcript view
6	Kozak sequence	Quality control	'Transcript Information' tab in Transcript view
7	Anti-sense gene information	Gene structure	'Gene structure' tab in Locus view
8	Detailed data of similarity to known ncRNA.	ncRNA	'Function' tab in Transcript view
9	Two new species (horse and medaka) for comparative analysis	Comparative	'Evolution' tab in Transcript view, G-integra and Evola
10	Detailed annotation for unmapped (UM) transcripts	Gene structure	Topic Annotation viewer
11	Remote integration of GlycoGene Database (GGDB)	Function	'Function' tab in Transcript view
12	Remote integration of the functional RNA database (fRNAdb)	ncRNA	'Function' tab in Transcript view

The Transcript and Locus views also have links to related external public databases including DDBJ/EMBL/GenBank (4), RefSeq (12), UniProtKB (13), HGNC (14), GeneCards (15), InterPro (16), Ensembl (17), EntrezGene (18), CCDS (19), PubMed (20), dbSNP (21), GO (22), GTOP (23), OMIM (24) and MutationView (25).

New features in G-integra

G-integra is an integrated genome browser in which we can examine the genomic structures of transcripts. The genomic locations, gene structures and alignments against the human genome of H-Inv transcripts, and the corresponding RefSeq and Ensembl entries are shown. We now show the annotations for two types of high-quality gene models, pHIT and eHIT, for all human gene tracks (Figure 1). G-integra provides gene structure annotations for two new species (horse and medaka). In total, the gene structures for humans and 13 non-human species, namely *Pan troglodytes* (chimpanzee), *Macaca* sp. (macaque), *Mus musculus* (mouse), *Rattus norvegicus* (rat), *Canis familiaris* (dog), *Bos taurus* (cow), *Monodelphis domestica* (opossum), *Gallus gallus* (chicken), *Equus ferus caballus* (horse), *Danio rerio* (zebrafish), *Tetraodon nigroviridis* (tetraodon), *Takifugu rubripes* (fugu) and *Oryzias latipes* (medaka) can be optionally displayed for comparison. The reference gene structures of non-coding RNAs of fRNAdb, pseudogenes of Pseudogene.org (26) and consensus coding sequences of CCDS (19) are also shown.

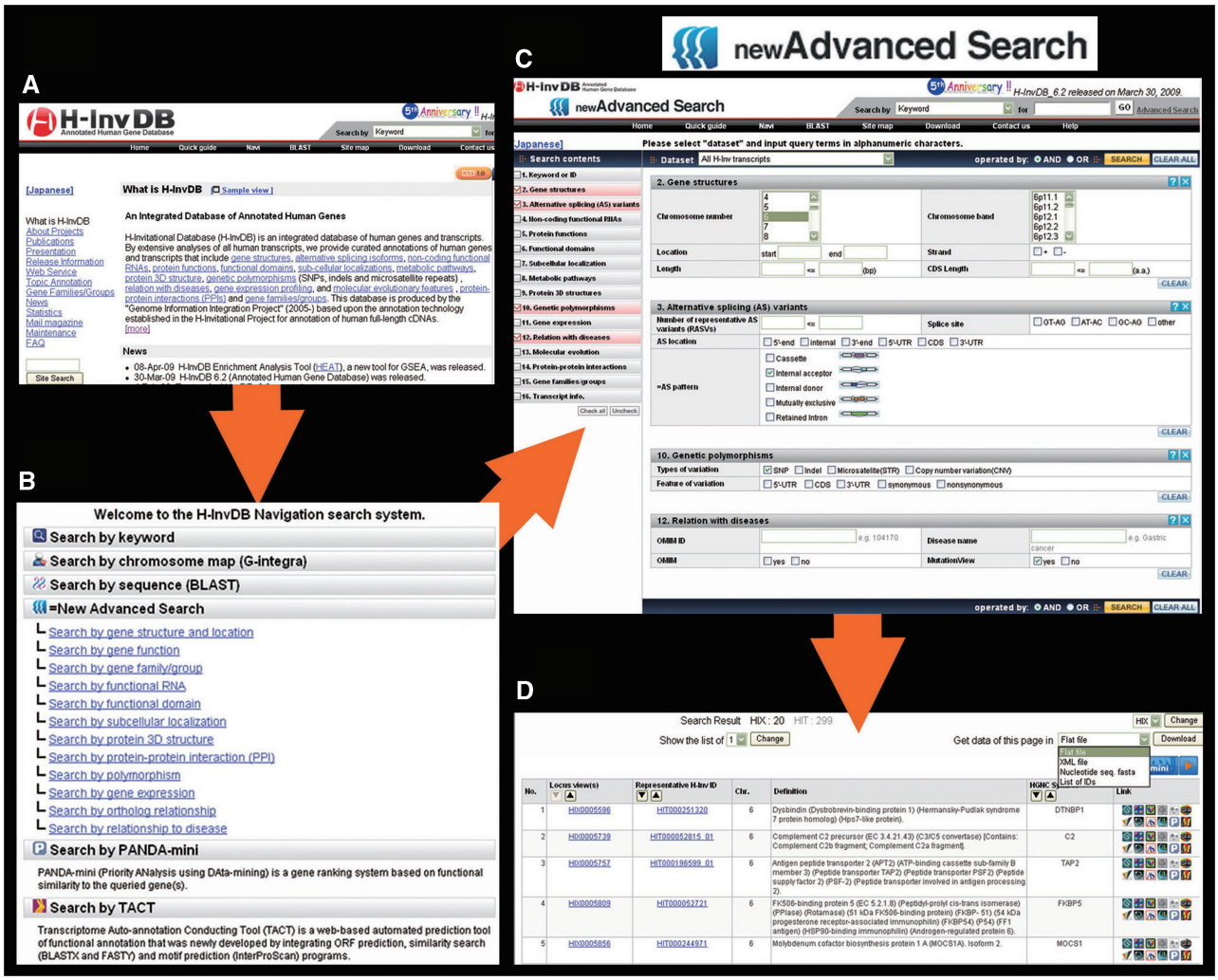

Figure 2. 'Navigation search': powerful search tool of 16 search items. Example screen shot of the Navigation search system (http://www.h-invitational.jp/hinv/c-search/). (**A**) There are links to the Navigation system, 'Navi', at the black menu bar in all the viewers in H-InvDB including the top page. (**B**) Search navigation menu provide the list of all searches available in H-InvDB. (**C**) The new advanced search provide combination search of 16 search contents, for example, #2 gene structure, #3 alternative splicing (AS) variants, #10 genetic polymorphism and #13 relation to disease. (**D**) The search results provide the list of HIX IDs, HIT IDs, Chromosome number, definition, HGNC gene symbol, and links to appropriate H-InvDB and related viewers.

NEWLY RELEASED DATA MINING RESOURCES IN H-InvDB

H-InvDB now provides newly released useful data mining resources, namely 'Navigation search', 'H-InvDB Enrichment Analysis Tool (HEAT)' and web service APIs.

Navigation search

'Navigation search' is an extended search system that enables complicated searches by any combination of 16 different search contents. This system consists of three interfaces: search navigation menu, new advanced search and search results and the user interface images are shown in Figure 2. Search navigation menu: for every view in H-InvDB for example the top page, there is a link to 'Navi' on the black menu bar (Figure 2A). The search navigation menu provides a list of all searches in H-InvDB (Figure 2B). New advanced search provides combined search of 16 search contents (Figure 2C). The search contents and items as summarized in Table 4. The search results page provides the search results and facilities to download the search results in four formats: flat file format, XML format, list of IDs in text format and sequence FASTA file (Figure 2D).

'Navigation search' provides the extended application for data mining of H-InvDB. For example, a user can search human genes for chromosome 6 with alternative splicing variants of an internal acceptor pattern, which contains an SNP and has disease information in OMIM (Figure 2C). To search new gene models, pHIT or eHIT transcripts, mol_type = predicted transcript (pHIT) or predicted transcript (eHIT) must be selected in the search content 'Transcript information'.

URL: http://h-invitational.jp/hinv/c-search/hinvNaviTop.jsp

Table 4. The list of search contents and items H-InvDB Navigation search

No.	Search content	Search items
1	Keyword or ID	13 IDs and 7 different types of keywords
2	Gene structure	chromosome number, chromosomal band, genome strand and location on the human genome
3	Alternative splicing (AS) variants	splicing site, pattern and location of alternative splicing
4	Non-coding functional RNAs	type and classification of ncRNAs
5	Protein functions	definition, similarity category, gene symbol, EC name and molecular function of GO
6	Functional domains	ID, name and type of InterPro domain
7	Subcellular localization	cellular component of GO and predicted subcellular localization by WoLF PSORT, SOSUI, TMHMM, TargetP and PTS1
8	Metabolic pathways	biological process of GO, ID and name of the KEGG pathway
9	Protein 3D structure	PDB and SCOP IDs of GTOP prediction
10	Genetic polymorphism	types and features of variation such as SNP, microsatellite, copy number variation (CNV), synonymous or nonsynonymous variations
11	Gene expression	tissue specific expression in ten tissue/organ classes, Affimetrix probe ID, promoter motif and upstream transcriptional start site (TSS)
12	Relation to disease	relation to MutationView, ID and disease name of OMIM
13	Molecular evolution	orthologues and genome conservation among human and 13 model organisms
14	Protein–protein interaction	number of interacting proteins
15	Gene families and groups	all the predicted human gene families and four manually curated gene families/groups; Ig, MHC, TCR and OR
16	Transcript information	sequence data provider, molecular type, coding potential and curation status information

H-InvDB Enrichment Analysis Tool

H-InvDB Enrichment Analysis Tool (HEAT) is a data mining tool for automatically identifying features specific to a given human gene set. HEAT searches for H-InvDB annotations that are significantly enriched in a user-defined gene set as compared with the entire H-InvDB representative transcripts. This technique is called 'gene set enrichment analysis' and is popularly used for analysing the results of microarray experiments. The HEAT analysis requires three steps. (i) Gene-Set Submission: users must submit two or more human gene IDs. Acceptable IDs are H-InvDB Transcript IDs (HIT), Locus IDs (HIX), HUGO Gene Symbols, and accession numbers of INSD (DDBJ/EMBL/GenBank). (ii) Execution: the submitted IDs are converted into HIXs of H-InvDB release 6.0 representative transcripts by using the ID Converter System (27). (iii) Results: enriched features of the given gene set are shown. For each feature, the link to description of the feature, number of occurrences/genes of a submitted gene set, number of occurrences/genes among all H-InvDB representative transcripts and *P*-values are shown. Features with *P*-values smaller than 0.01 are shown and the list of results are sorted by *P*-value. Fisher's exact probability is used in calculating the *P*-values. The following features of H-InvDB are analysed: InterPro, GO, the KEGG pathway, chromosomal band, gene family, structural domains (SCOP), subcellular localization prediction (using WoLF PSORT) and tissue-specific gene expression (10 tissue categories defined in H-ANGEL).

URL: http://hinv.jp/HEAT/search.php?lang=en.

H-InvDB web-service APIs: a new data retrieval service

The web service interface is becoming a major way for accessing biological databases (28). H-InvDB now provides a new data retrieval service, web service with APIs of Simple Object Access Protocol (SOAP) and Representational State Transfer (REST), to retrieve the H-InvDB entries of given IDs or keywords. Entries in H-InvDB can be retrieved in XML or sequence FASTA format. The current H-InvDB web service provides 26 SOAP and 28 REST APIs. To use the REST service, an HTTP connection (e.g. web browser) and a programming language (e.g. Perl, JAVA) are required. Although both the POST and GET methods of access are supported, the POST method is approved. To retrieve entries for a keyword, e.g. 'cancer', the method and parameters are as follows: http://h-invitational.jp/hinv/hws/keyword_search.php?query=cancer.

To use the SOAP service, users are requested to use the SOAP library of programming languages. Access to WSDL is via http://h-invitational.jp/hinv/hws/API?wsdl. The 12 representative SOAP APIs are listed in Table 5, and complete detailed descriptions are provided at the following URLs:

REST APIs: http://www.h-invitational.jp/hinv/hws/doc/en/api_list.php

SOAP APIs: http://www.h-invitational.jp/hinv/hws/doc/en/soap_api_list.php

The H-InvDB web service is already used for retrieving H-InvDB data by other databases. For example, in MutationView, a database for mutations in human disease genes (25), the InterPro domain data in H-InvDB are used to search for relations among of the functional domains, human genes and human disease-related mutations.

DATA AVAILABILITY AND FUTURE DIRECTIONS

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, original-format flat files or XML files at HTTP and FTP servers. Major

Table 5. The list of representative H-InvDB web service APIs (SOAP)

API type	Description of API	WDSL	Query and output
Search entries	Search by IDs	soap_id_search.php?wsdl	query = any ID output = HIT ID
	Search by keywords	soap_keyword_search.php?wsdl	query = any keyword output = HIT ID
	Search by genomic location	soap_location2hit.php?wsdl	query = genomic location output = corresponding HIT ID
Count entries	Total number of HIT	soap_hit_cnt.php?wsdl	output = total number of HIT ID
Convert IDs	Convert ISND accession to HIT	soap_acc2hit.php?wsdl	query = Accession No. output = HIT ID
Retrieve data	Retrieve HIT XML file	soap_hit_xml.php?wsdl	query = HIT ID output = HIT XML file
	Retrieve HIT definition	soap_hit_definition.php?wsdl	query = HIT ID output = HIT definition
	Retrieve HIT evolutionary information	soap_hit_evolution.php?wsdl	query = HIT ID output = evolutionary information
	Retrieve HIT gene expression information	soap_hit_expression.php?wsdl	query = HIT ID output = gene expression information
	Retrieve HIT genomic location of HIT	soap_hit_location.php?wsdl	query = HIT ID output = genomic location of HIT
	Retrieve nucleotide sequence of HIT	soap_hit_nucleotide_seq_xml.php?wsdl	query = HIT ID output = nucleotide sequence of HIT (XML format)
	Retrieve protein sequence of HIT	soap_hit_protein_seq_xml.php?wsdl	query = HIT ID output = protein sequence of HIT (XML format)

updates are released once a year and minor updates are released a few times per year when necessary. For the next major update of H-InvDB by the end of this year, the annotations for the latest human genome assembly NCBI b37 will be provided.

ACKNOWLEDGEMENTS

The authors acknowledge all the members of the H-Invitational consortium and the Genome Information Integration Project (GIIP) for participating in the annotation work of human full-length cDNAs and all the staffs of the Integrated Database and Systems Biology Team of BIRC, AIST, for supporting the construction of H-InvDB. We thank Dr. Satoshi Fukuchi of National Institute of Genetics, Dr. Paul Horton of Computational Biology Research Center, National Institute of Advanced Industrial Science and Technology, and Dr. Mitsuteru Nakao of Kazusa DNA Research Institute for their special cooperation to H-InvDB annotation.

FUNDING

Ministry of Economy, Trade and Industry of Japan (METI); the National Institute of Advanced Industrial Science and Technology (AIST); the Japan Biological Informatics Consortium (JBIC). Funding for open access charge: Advanced Industrial Science and Technology.

Conflict of interest statement. None declared.

REFERENCES

1. Imanishi,T., Itoh,T., Suzuki,Y., O'Donovan,C., Fukuchi,S., Koyanagi,K.O., Barrero,R.A., Tamura,T., Yamaguchi-Kabata,Y., Tanino,M. *et al.* (2004) Integrative annotation of 21,037 human genes validated by full-length cDNA clones. *PLoS Biol.*, **2**, 856–875.
2. Yamasaki,C., Koyanagi,K., Fujii,Y., Itoh,T., Barrero,R., Tamura,T., Yamaguchi-Kabata,Y., Tanino,M., Takeda,J., Fukuchi,S. *et al.* (2005) Investigation of protein functions through data-mining on integrated human transcriptome database, H-Invitational database (H-InvDB). *Gene*, **364**, 99–107.
3. Yamasaki,C., Murakami,K., Fujii,Y., Sato,Y., Harada,E., Takeda,J., Taniya,T., Sakate,R., Kikugawa,S., Shimada,M. *et al.* (2008) The H-Invitational Database (H-InvDB), a comprehensive annotation resource for human genes and transcripts. *Nucleic Acids Res.*, **36**, D793–D799.
4. Tateno,Y. (2008) International collaboration among DDBJ, EMBL Bank and GenBank. *Tanpakushitsu Kakusan Koso*, **53**, 182–189.
5. Allen,J.E. and Salzberg,S.L. (2005) JIGSAW: integration of multiple sources of evidence for gene prediction. *Bioinformatics*, **21**, 3596–3603.
6. Tanino,M., Debily,M.A., Tamura,T., Hishiki,T., Ogasawara,O., Murakawa,K., Kawamoto,S., Itoh,K., Watanabe,S., de Souza,S.J. *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**, D567–D572.
7. Matsuya,A., Sakate,R., Kawahara,Y., Koyanagi,K.O., Sato,Y., Fujii,Y., Yamasaki,C., Habara,T., Nakaoka,H., Todokoro,F. *et al.* (2008) Evola: Ortholog database of all human genes in H-InvDB with manual curation of phylogenetic trees. *Nucleic Acids Res.*, **36**, D787–D792.
8. Takeda,J., Suzuki,Y., Sakate,R., Sato,Y., Seki,M., Irie,T., Takeuchi,N., Ueda,T., Nakao,M., Sugano,S. *et al.* (2008) Low conservation and species-specific evolution of alternative splicing in humans and mice: comparative genomics analysis using well-annotated full-length cDNAs. *Nucleic Acids Res.*, **36**, 6386–6395.
9. Kozak,M. (1984) Compilation and analysis of sequences upstream from the translational start site in eukaryotic mRNAs. *Nucleic Acids Res.*, **12**, 857–872.
10. Narimatsu,H. (2004) Construction of a human glycogene library and comprehensive functional analysis. *Glycoconj J.*, **21**, 17–24.
11. Kin,T., Yamada,K., Terai,G., Okida,H., Yoshinari,Y., Ono,Y., Kojima,A., Kimura,Y., Komori,T. and Asai,K. (2007) fRNAdb: a platform for mining/annotating functional RNA candidates from non-coding RNA sequences. *Nucleic Acids Res.*, **35**, D145–D148.
12. Pruitt,K.D., Tatusova,T. and Maglott,D.R. (2007) NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.*, **35**, D61–D65.
13. UniProt Consortium. (2009) The Universal Protein Resource (UniProt) 2009. *Nucleic Acids Res.*, **37**, D169–D174.
14. Bruford,E.A., Lush,M.J., Wright,M.W., Sneddon,T.P., Povey,S. and Birney,E. (2008) The HGNC Database in 2008: a resource for the human genome. *Nucleic Acids Res.*, **36**, D445–D448.
15. Safran,M., Chalifa-Caspi,V., Shmueli,O., Olender,T., Lapidot,M., Rosen,N., Shmoish,M., Peter,Y., Glusman,G., Feldmesser,E. *et al.* (2003) Human Gene-Centric Databases at the Weizmann Institute

of Science: GeneCards, UDB, CroW 21 and HORDE. *Nucleic Acids Res.*, **31**, 142–146.
16. Hunter,S., Apweiler,R., Attwood,T.K., Bairoch,A., Bateman,A., Binns,D., Bork,P., Das,U., Daugherty,L., Duquenne,L. *et al.* (2009) InterPro: the integrative protein signature database. *Nucleic Acids Res.*, **37**, D211–D215.
17. Hubbard,T.J., Aken,B.L., Ayling,S., Ballester,B., Beal,K., Bragin,E., Brent,S., Chen,Y., Clapham,P., Clarke,L. *et al.* (2009) Ensembl 2009. *Nucleic Acids Res.*, **37**, D690–D697.
18. Maglott,D., Ostell,J., Pruitt,K.D. and Tatusova,T. (2007) Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res.*, **35**, D26–D31.
19. Pruitt,K.D., Harrow,J., Harte,R.A., Wallin,C., Diekhans,M., Maglott,D.R., Searle,S., Farrell,C.M., Loveland,J.E., Ruef,B.J. *et al.* (2009) The consensus coding sequence (CCDS) project: Identifying a common protein-coding gene set for the human and mouse genomes. *Genome Res.*, **19**, 1316–1323.
20. Giglia,E. (2009) Medline/PubMed revisited: new, semantic tools to explore the biomedical literature. *Eur. J. Phys. Rehabil. Med.*, **45**, 293–297.
21. Sherry,S.T., Ward,M.H., Kholodov,M., Baker,J., Phan,L., Smigielski,E.M. and Sirotkin,K. (2001) dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.*, **29**, 308–311.
22. Barrell,D., Dimmer,E., Huntley,R.P., Binns,D., O'Donovan,C. and Apweiler,R. (2009) The GOA database in 2009 – an integrated Gene Ontology Annotation resource. *Nucleic Acids Res.*, **37**, D396–D403.
23. Fukuchi,S., Homma,K., Sakamoto,S., Sugawara,H., Tateno,Y., Gojobori,T. and Nishikawa,K. (2009) The GTOP database in 2009: updated content and novel features to expand and deepen insights into protein structures and functions. *Nucleic Acids Res.*, **37**, D333–D337.
24. Amberger,J., Bocchini,C.A., Scott,A.F. and Hamosh,A. (2009) McKusick's Online Mendelian Inheritance in Man (OMIM). *Nucleic Acids Res.*, **37**, D793–D796.
25. Shimizu,N., Ohtsubo,M. and Minoshima,S. (2007) MutationView/KMcancerDB: a database for cancer gene mutations. *Cancer Sci.*, **98**, 259–267.
26. Karro,J.E., Yan,Y., Zheng,D., Zhang,Z., Carriero,N., Cayting,P., Harrrison,P. and Gerstein,M. (2007) Pseudogene.org: a comprehensive database and comparison platform for pseudogene annotation. *Nucleic Acids Res.*, **35**, D55–D60.
27. Imanishi,T. and Nakaoka,H. (2009) Hyperlink Management System and ID Converter System: enabling maintenance-free hyperlinks among major biological databases. *Nucleic Acids Res.*, **37**, W17–W22.
28. McWilliam,H., Valentin,F., Goujon,M., Li,W., Narayanasamy,M., Martin,J., Miyar,T. and Lopez,R. (2009) Web services at the European Bioinformatics Institute-2009. *Nucleic Acids Res.*, **37**, W6–W10.