Nature Precedings : doi:10.1038/npre.2010.5271.1 : Posted 17 Nov 2010

H-InvDB, A Comprehensive Annotation Resource For Human Transcriptome

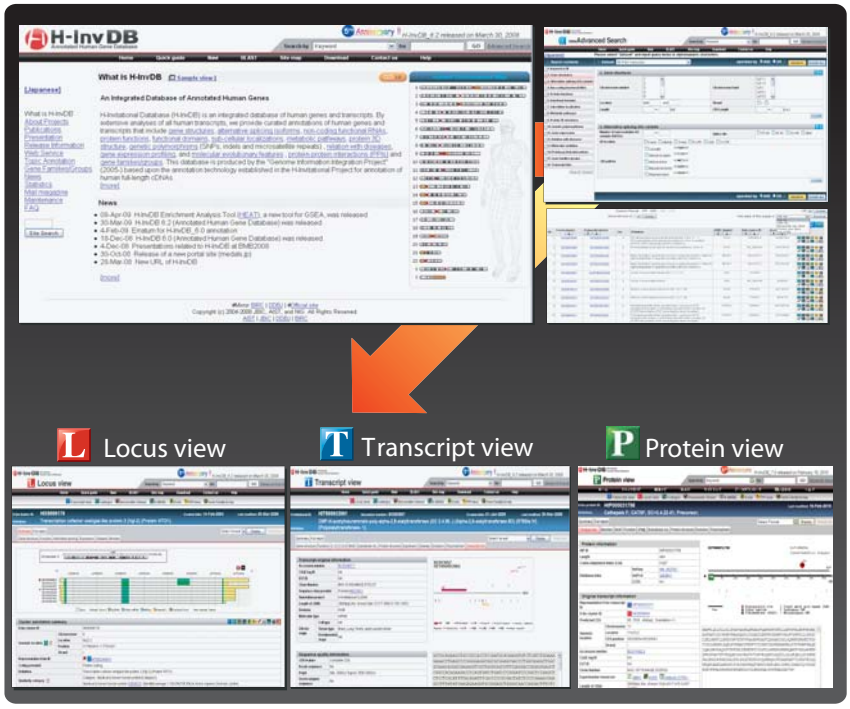
○Chisato Yamasaki (1), Jun-ichi Takeda (1), Takuya Habara (1), Makoto Ogawa (2,3), Akiko Noda (1), Ryuichi Sakate (1), Katsuhiko Murakami (2), Tadashi Imanishi (1), Takashi Gojobori (1, 4)
1: BIRC, AIST, 2: JBIC, 3: DYNACOM Co., Ltd., 4: CIB-DDBJ, NIG

H-Invitational Database (H-InvDB; http://www.h-invitational.jp/) is a comprehensive annotation resource for human transcriptome.

By extensive analyses of all human transcripts, we provide curated annotations of human genes, transcripts and proteins that include gene structures, alternative splicing isoforms, non-coding functional RNAs, protein functions, functional domains, sub-cellular localizations, metabolic pathways, protein 3D structure, genetic polymorphisms, relation with diseases, gene expression profiling, molecular evolutionary features, protein-protein interactions (PPIs) and gene families/groups.

The latest release of H-InvDB (release 7.5) provides annotation for 242,813 human transcripts in 44,806 human gene clusters based on human full-length cDNAs, mRNAs and the reference human genome sequences (NCBI b37.1). H-InvDB consists of three main views, the Transcript view, the Locus view and the Protein view, and six sub-databases; G-integra, H-ANGEL, DiseaseInfo Viewer, Evola, PPI view and Gene Family/Group view. We also provide data mining tools such as "Navigation search" , an extended search system that enables complicated searches by combining 16 different search options (http://www.h-invitational.jp/hinv/c-search/hinvNaviTop.jsp) and "H-InvDB Enrichment Analysis Tool (HEAT)" , a data mining tool for automatically identifying features specific to a given human gene set (http://hinv.jp/HEAT/).

H-InvDB databases



Locus view

H-InvDB provides annotation items for each HIX (H-Invitational cluster) in Locus view

10. **Locus info.** **Locus information** mapping of HIX (H-Invitational cluster) in the human genome; chromosome number, location, strand, chromosome band, disease relationship and links to corresponding RefSeq and Ensembl genes, etc.
11. **AS** **Alternative splicing (AS) information** annotation on alternative splicing isoforms.
12. **Expression** **Gene expression information** tissue-specific expression in 10 tissue categories determined by iAFLP data.
13. **Disease info** **Disease/pathology information** disease related information related to HIX: known disease-related genes and co-localized orphan pathology with the name of the disease and OMIM ID.

Transcript view

H-InvDB provides annotation items for each HIT (H-Invitational transcript) in Transcript view

1. **Function** **Gene function information** human-curated functional definition, similarity category and related evidences; Gene name; HUGO gene symbols; GO ID; GO term; EC number; EC description; pathway information (KEGG), etc
2. **Genome loc.** **Genome location information** mapping of HIT on the human genome; chromosome number, location, strand, chromosome band, and links to corresponding RefSeq and Ensembl genes, etc
3. **Transcript info.** **Transcript information/Transcript quality information** transcript length, polyA signal, polyA tail and sequence quality related features.
4. **Polymorphism/repeat** **Polymorphism (SNP, indel), microsatellite (Short Tandem Repeat, STR) and repeat information** polymorphism (dbSNP), Microsatellite (Short Tandem Repeat, STR) and repeat information.
5. **CDS** **Predicted CDS information** CDS, orientation, codon adaptation index, translation.
6. **Evolutionary info.** **Evolutionary information** Ortholog relationships, phylogenic trees and sequence alignments.

Protein view

H-InvDB provides annotation items for each HITP(H-Invitational protein in Protein view

1. **Protein info.** **Protein information**
2. **Function** **Gene function information** human-curated functional definition, similarity category and related evidences; Gene name; HUGO gene symbols; GO ID; GO term; EC number; EC description; pathway information (KEGG), etc
3. **Motif** **Motif information** location, ID and descriptions of functional motifs (InterPro).
4. **Subcellular loc.** **Subcellular localization information** subcellular localization prediction by WoIfPSORT Target P, SOSUI, TMHMM and PTS1.
5. **Protein structure** **Protein structure information (G TOP)** assigned PDB and SCO PIDs by reverse PSI-BLAST and summary prediction of 3D structure by GTO P

H-InvDB sub-databases

G-integra

G-integra is an original genome browser, in which we can browse physical maps and gene structures of human and 13 model organisms (mouse, rat, chimpanzee, orangutan, rhesus monkeychicken, dog, horse, cow, opossum, hugu, tetaodon, medaka, zebrafish)

H-ANGEL

H-ANGEL (Human ANatomic Gene Expression Library) is a database of expression profiles of human genes. Gene expression data in normal adult human tissues that were generated by three types of methods and in seven different platforms were collected and categorized into 10 and 40 major tissues.

DiseaseInfo Viewer

DiseaseInfo Viewer is a database of known and orphan genetic diseases and their relation to H-Inv loci with OMIM and MutationView.

Evola

Evola contains ortholog information among human and other vertebrates based on our analyses of comparative genomics and transcriptomics. Evola displays sequence alignments, phylogenetic trees and gene loci of orthologs.

PPI view

The PPI view displays H-InvDB human protein-protein interaction (PPI) information. It integrated PPI information of BIND, DIP, MINT, HPRD, IntAct and GNP_Y2H databases and mapped on to H-Inv proteins (HIP).

Gene family / group view

H-InvDB Gene Families/Groups are human-curated annotation datasets of a selected gene families/groups (ie. TCR, MHC, OR) . It also provide the predicted human gene families/groups based on the sequence similarities and assigned an HIF ID for each H-Inv gene family/group.

H-InvDB satellite-databases and tools

- LEGENDA :Literature-Extracted GENe-Disease Associations
  A database that was built based on the extraction of the relations between genes, diseases, and substances from the entire MEDLINE titles and abstracts.
- H-GOLD :Human-Gene diversity Of Life-style related Disease
  A database of ca. 30,000 Microsatellite markers, which are detected in silico from genome sequence and confirmed polymorphism in Japanese population
- G-compass :Comparative genome browser
  A web tool for comparative genomics, which can browse conserved regions based on the genome alignments of human-chimpanzee, human-mouse and human-rat.
- TACT :Transcriptome Auto-annotation Conducting Tool
  A web-based automated prediction tool of functional annotation that was newly developed by integrating ORF prediction, similarity searches and motif prediction programs
- H-DBAS :Human transcriptome DataBase for Alternative Splicing
  The representative AS variants (RASV) were selected among the cluster of AS variants which have the same genomic (exon-intron) structure.
- VarySysDB :Database of annotated human polymorphism
  It offers annotated human polymorphism information of single nucleotide polymorphisms (SNPs) on splice sites and transcripts, deletion-insertion polymorphisms (DIPs), short tandem repeats (STRs), single amino acid repeats (SARs), structural variation (or copy number variations: CNVs), linkage disequilibrium regions, and their relations to the genome, transcripts, and functional domains.

http://www.h-invitational.jp/ [hinv.jp]

H-InvDB Enrichment Analysis Tool (HEAT)

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

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. The following features of H-InvDB are analyzed: InterPro, Gene Ontology (GO), KEGG pathway, Chromosomal band, Gene family, SCOP (structural domains), Subcellular localization prediction (by using Wolf PSORT) and Tissue-specific gene expression (10 tissue categories defined in H-ANGEL). This technique is called Gene Set Enrichment Analysis (GSEA), and is popularly used in analyzing results of microarray experiments..

References:

(1) H-InvDB in 2009, extended database and data mining resources for human genes and transcripts. Yamasaki C, et al. (2010) Nucleic Acids Research 38 (Database Issue) (in press).
(2) The H-Invitational Database (H-InvDB), a comprehensive annotation resource for human genes and transcripts. Yamasaki C, et al. (2008) Nucleic Acids Research 36, Database issue D793-D799.
(3) Integrative Annotation of 21,037 Human Genes Validated by Full-Length cDNA Clones. T. Imanishi et al. (2004) PLoS Biology 2 (6), 856-875.

Biomedicinal Information Research Center (BIRC), National Institute of Advanced Industrial Science and Technology (AIST)
2-4-7 Aomi, Koto-ku, Tokyo 135-0064, Japan. URL http://www.h-invitational.jp/idb/index_e.html