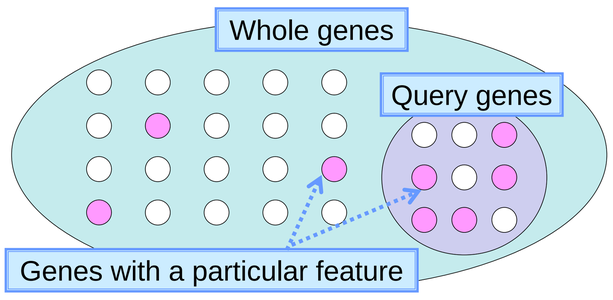
HEAT: a New Tool For Gene Set Enrichment Analysis Using Comprehensive Annotation of Human Genes in H-InvDB

Tadashi Imanishi, Akiko Ogura Noda, and Miho Sera

Biomedicinal Information Research Center, National Institute of Advanced Industrial Science and Technology, JAPAN

H-InvDB Enrichment Analysis Tool (HEAT) is a new data-mining tool for gene set enrichment analysis based on comprehensive annotations of human genes in H-InvDB. 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 advantage of HEAT is the wide variety of annotation items used for its analysis: chromosomal bands, InterPro functional domains, Gene Ontology terms, KEGG pathways, H-InvDB gene families/groups, SCOP structural domains, subcellular localization predicted by using the Wolf-PSORT program, tissue-specific gene expression as defined in the H-ANGEL database, and transcription factor binding sites in promoter regions based on JASPAR. HEAT accepts lists of human gene identifiers (IDs) including HUGO gene symbols, accession numbers of INSD (DDBJ/EMBL/GenBank), UniProt accession numbers, Gene IDs, Ensembl Gene IDs, H-InvDB Transcript IDs (HIT) and Locus IDs (HIX), etc. Then, HEAT converts the accepted IDs into HIX using the ID Converter System (http://biodb.jp/), collects various annotations of H-InvDB representative transcripts, and conducts statistical tests by using Fisher's exact probability. The output of HEAT is a simple report of annotations commonly found among the query genes, which is very useful to grasp the property of a particular gene set. HEAT is freely available at http://hinv.jp/HEAT/search.php?lang=en.

Introduction



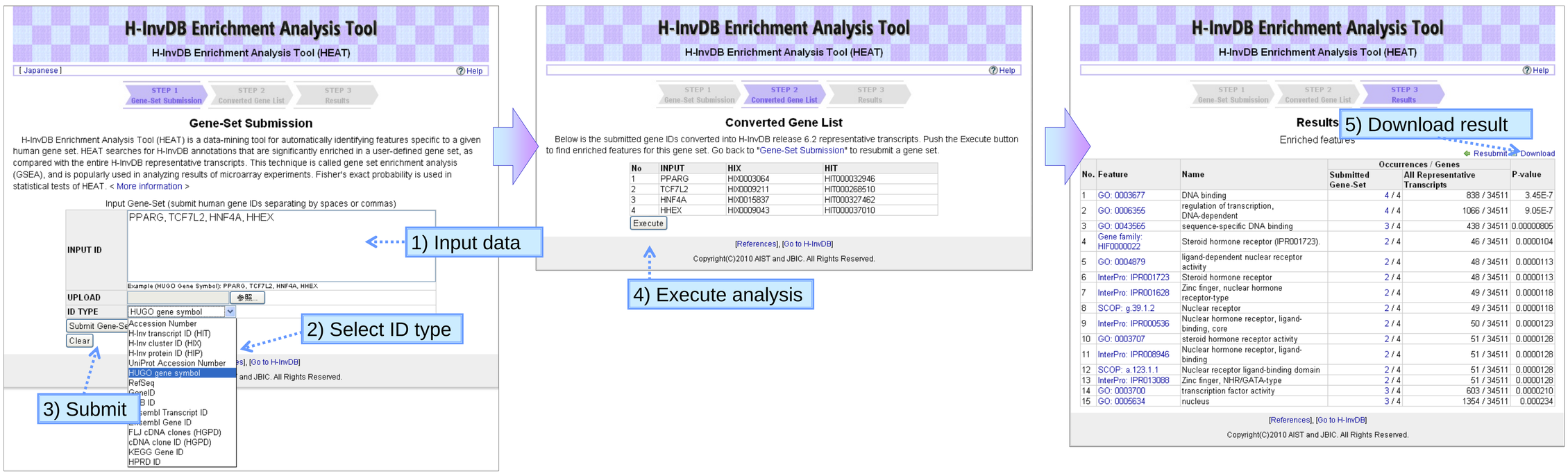
Gene set enrichment analysis is a method to find features enriched in a given gene set. The following 10 annotations are used for analysis in HEAT.

- InterPro
- Gene Ontology
- KEGG pathway
- Chromosomal band
- Gene family
- SCOP (Structural domains)
- Subcellular localization (predicted by Wolf PSORT)
- Subcellular localization (predicted by SOSUI)
- Tissue-specific gene expression (H-Angel)
- Sequence motifs in promoter regions (JASPAR)

HEAT conducts gene set enrichment analysis by calculating P-values using Fisher's exact probability test. Let N be the number of all protein-coding representative transcripts in H-InvDB, n be the number of occurrences of a particular feature, K be the number of genes submitted, and k be the number of occurrences of the feature in the given gene set. Then, the P-value is calculated by the following equation.

$$P = \sum_{i \geq k} \frac{{}_nC_i \times {}_{(N-n)}C_{(K-i)}}{{}_NC_K}$$

Usage



Input gene list. HEAT accepts 15 types of public IDs.

Check the converted gene list. HEAT automatically converts submitted IDs into H-invDB locus IDs (HIX), in order to utilize the annotation resource of H-invDB.

Get the result. Enriched features are listed in ascending order of P-value. The result list is also provided in a downloadable text file.

Application

A list of 3,947 genes that show differential expression profile between ES cells and early passage iPS cells (Ref.4, supplementary data) is analyzed with HEAT. The top 20 resultant features are shown.

No.	feature		occurrence in query set	occurrence in background set	P-value
1	JASPAR: MA0003.1	TFAP2A	2096 / 3446	12567 / 34511	5.13E-207
2	JASPAR: MA0079.2	SP1	2676 / 3446	18952 / 34511	1.51E-187
3	JASPAR: MA0117.1	Mafb	1360 / 3446	8549 / 34511	3.01E-90
4	JASPAR: MA0028.1	ELK1	945 / 3446	5243 / 34511	2.24E-85
5	JASPAR: MA0006.1	Arnt::Ahr	1114 / 3446	7099 / 34511	7.66E-66
6	JASPAR: MA0146.1	Zfx	1884 / 3446	14187 / 34511	2.79E-64
7	JASPAR: MA0039.2	Klf4	1936 / 3446	14752 / 34511	8.85E-63
8	JASPAR: MA0062.2	GABPA	802 / 3446	4785 / 34511	4.58E-56
9	JASPAR: MA0104.2	Mycn	1206 / 3446	8203 / 34511	1.12E-55
10	JASPAR: MA0259.1	HIF1A::ARNT	1454 / 3446	10729 / 34511	7.67E-48
11	GO: 0003676	nucleic acid binding	307 / 3446	1336 / 34511	8.54E-46
12	GO: 0005622	intracellular	356 / 3446	1785 / 34511	1.15E-38
13	JASPAR: MA0162.1	Egr1	760 / 3446	4941 / 34511	1.85E-38
14	JASPAR: MA0147.1	Myc	1053 / 3446	7539 / 34511	1.62E-36
15	GO: 0008270	zinc ion binding	330 / 3446	1656 / 34511	9.82E-36
16	JASPAR: MA0014.1	Pax5	688 / 3446	4523 / 34511	7.20E-33
17	GO: 0005515	protein binding	271 / 3446	1327 / 34511	3.74E-31
18	JASPAR: MA0004.1	Arnt	1014 / 3446	7442 / 34511	1.70E-30
19	GO: 0005634	nucleus	270 / 3446	1354 / 34511	2.65E-29
20	JASPAR: MA0056.1	MZF1_1-4	1247 / 3446	9663 / 34511	1.69E-28

- Result:742 annotations
  - CytoBand 11
  - Gene family 85
  - Gene Ontology 129
  - InterPro 279
  - JASPAR 44
  - KEGG 50
  - SCOP 135
  - SOSUI 1
  - Tissue-specificity 2
  - WPSORT 6

[REFERENCES]

1. Yamasaki C, Murakami K, Takeda J, Sato Y, Noda A, Sakate R, Habara T, Nakaoka H, Todokoro F, Matsuya A, Imanishi T, and Gojobori T (2009) H-InvDB in 2009, extended database and data mining resources for human genes and transcripts. *Nucleic Acids Research* **38** (Database Issue): D626-D632.
2. Imanishi T, Itoh T, Suzuki Y, O'Donovan C, Fukuchi S, Koyanagi KO, Barrero RA, Tamura T, Yamaguchi-Kabata Y, Tanino M, *et al.* (2004) Integrative annotation of 21,037 human genes validated by full-length cDNA clones. *PLoS Biology* **2:** 856-875.
3. Imanishi T and Nakaoka H (2009) Hyperlink Management System and ID Converter System: enabling maintenance-free hyperlinks among major biological databases. *Nucleic Acids Research* **37** (Web Server Issue): W17-W22 (gkp355).
4. Chin MH, *et al.* (2009) Induced pluripotent stem cells and embryonic stem cells are distinguished by gene expression signatures. Cell Stem Cell **5**:111-23.

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