W152–W158 *Nucleic Acids Research, 2007, Vol. 35, Web Server issue*
doi:10.1093/nar/gkm381

GeneHub-GEPIS: digital expression profiling for normal and cancer tissues based on an integrated gene database

Yan Zhang[1], Shiuh-Ming Luoh[1], Lawrence S. Hon[1], Robert Baertsch[2], William I. Wood[1] and Zemin Zhang[1],*

[1]Department of Bioinformatics, Genentech, Inc., South San Francisco, CA 94080, USA and
[2]Department of Biomolecular Engineering, University of California, Santa Cruz, CA 95064, USA

Received January 31, 2007; Revised April 27, 2007; Accepted April 30, 2007

ABSTRACT

GeneHub-GEPIS is a web application that performs digital expression analysis in human and mouse tissues based on an integrated gene database. Using aggregated expressed sequence tag (EST) library information and EST counts, the application calculates the normalized gene expression levels across a large panel of normal and tumor tissues, thus providing rapid expression profiling for a given gene. The backend GeneHub component of the application contains pre-defined gene structures derived from mRNA transcript sequences from major databases and includes extensive cross references for commonly used gene identifiers. ESTs are then linked to genes based on their precise genomic locations as determined by GMAP. This genome-based approach reduces incorrect matches between ESTs and genes, thus minimizing the noise seen with previous tools. In addition, the gene-centric design makes it possible to add several important features, including text searching capabilities, the ability to accept diverse input values, expression analysis for microRNAs, basic gene annotation, batch analysis and linking between mouse and human genes. GeneHub-GEPIS is available at http://www.cgl.ucsf.edu/Research/genentech/genehub-gepis/ or http://www.gepis.org/.

INTRODUCTION

Analysis of gene expression profiles across various tissue types is essential for understanding gene functions. Among various available expression data sources, expressed sequence tags (ESTs) have been valuable for rapid expression profiling. Based on the premise that EST clone frequency is proportional to the corresponding gene's expression level (1), we and others have developed algorithms and tools to perform expression analysis based on EST data (2–7). Meanwhile, EST data continue to accumulate at a rapid pace, and there are a growing number of databases that organize general or species-specific EST information, including EST data for sea bass, wheat, chicken, pig and tomato (8–12). Despite the surge of recent progress in other species, the number of public EST entries for human and mouse still far exceed those for any other species, based on the January 2007 summary from dbEST (http://www.ncbi.nlm.nih.gov/dbEST/). Since the reliability of EST-based expression analysis is dependent on the size of EST libraries, the human and mouse data remain an attractive source for expression analysis, and the tools built for analyzing these data will likely benefit expression analysis for other species.

We previously developed the GEPIS server that utilizes EST abundance information to calculate gene expression levels in a panel of normal and cancerous human tissues for a given input DNA sequence (7). We showed that such EST-based (or 'digitally' derived) expression units exhibit a linear correlation with TaqMan-determined expression levels. Since its release, the GEPIS server has provided expression results for over 30 000 requests by researchers from >60 countries. Despite its usefulness, GEPIS suffers from several limitations. The method relied on the BLAST program to assign EST sequences to a given input mRNA sequence. However, BLAST often erroneously links ESTs to input sequences due to high-percentage regional matches. As a result, a given EST could be matched to multiple genes, thus leading to miscalculated expression data. In addition, there were insufficient data for performing reliable analysis for mouse genes. The design of the system also did not allow easy development of new functionality commonly requested by users, such as URL linking to the expression results, text searching and display of detailed results.

*To whom correspondence should be addressed. Tel: +1 650 225 4293; Fax: 650 225 5389; Email: zemin@gene.com

We have now developed a new web server, named GeneHub-GEPIS, which performs digital expression analysis based on an integrated database for human and mouse genes. One distinguishing characteristic of this tool is that ESTs are mapped to pre-defined gene structures along the genome. The GeneHub component of the application is designed to define gene boundaries based on mRNA transcript sequences from major databases and to establish extensive cross references for commonly used gene identifiers. Based on the precise genomic locations of ESTs, as determined by the GMAP algorithm (13), we link ESTs to genes for subsequent expression analysis. The new design offers several major advantages. First, this genome-based approach increases the accuracy of EST mapping to genes, thus enhancing the overall reliability of the EST-based expression values. Second, the new gene-centric design makes the system more extensible, so that we could easily add a new collection of genes, such as microRNAs, to the system. Third, the integrated gene database accepts text-based searches and diverse input values. In addition to DNA sequences, the input values can be identifiers from common gene databases or commercial microarray platforms. Fourth, the orthologous relationships stored in the GeneHub database allow easy navigation between mouse and human genes. Finally, the program provides basic information about input genes and allows direct linking to expression results from any web site. Meanwhile, we retained the useful features from the previous GEPIS application, such as the capability to draw a regional expression atlas for a given genomic region. Here, we present GeneHub-GEPIS as a new and useful tool for performing gene expression analysis across many normal and cancer tissues for both mouse and human genes.

MATERIALS AND METHODS

Genome-guided gene definition and cross references

The genomic structures of protein-coding genes were first defined using transcripts from several reliable sources. The collection of such high-quality transcripts, which we also call the core gene set, contains mRNA sequences from RefSeq, the Known gene set of Ensembl genes, Proteome and FANTOM (mouse only) (see Supplementary Table 1 for details). Each of the core gene set sequences was mapped to their respective genome (human NCBI Build 36 and mouse NCBI Build 35) using GMAP (13), and only the genomic match with the highest percent identity and percent coverage was chosen. GMAP has been shown to provide very accurate mapping and alignment results for both mRNA transcripts and ESTs (13), but occasionally matches with lower matching percentage can be found due to low-quality sequences. As a conservative precaution, we removed transcripts (and ESTs for a later step) with <90% coverage of the entire transcript or with <90% identity as measured by GMAP. For instance, about 0.4% RefSeq sequences were filtered out during this step. For any two transcripts to be clustered into one gene, we required that their exon sequences overlap, be in the same orientation and share at least one exact exon boundary or splice site. The requirement for shared exon boundary was used to limit the inclusion of antisense transcripts, since the orientation of these transcripts could be occasionally mis-annotated (14,15). Each group (or cluster) of transcripts was considered as a ‘GeneHub gene’. Using this approach, we defined 31 999 non-redundant genes for the human and 34 794 for mouse. Overall, these GeneHub genes can be considered as a superset of known genes from the above data sources.

After the initial collection of human and mouse genes were obtained, additional sequences were mapped to the GeneHub genes. Transcripts from GenBank and the Ensembl Novel collection were mapped to the genomes with GMAP (13) and then compared with those GeneHub genes derived above. For a transcript to be linked to a GeneHub gene, at least one of the splicing junctions was required to match perfectly with those of the GeneHub gene. We next tried to assign microarray-related probe sequences (see Supplementary Table 1 for the full list) from commonly used commercial array platforms to GeneHub genes. For Affymetrix expression arrays, we used the target sequences obtained from Affymetrix to link to known genes. For Agilent oligo-based expression microarrays, we directly used the 60-mer oligo-nucleotide sequences for gene linking. Using GMAP (13), we determined whether the array probe sequences overlap with the exon sequences collected above. Next, for sequences that did not overlap with any exons, we examined whether they were located in the vicinity of any GeneHub genes. We assigned a probe sequence to the closest gene in the same orientation if the probe sequence was located within 5 kb to the 3′ end or 2 kb to the 5′ end of the gene.

Additional methods were used for both human and mouse data to link other protein or DNA identifiers to GeneHub genes. First, for the roughly 1% of all transcript sequences that failed to align to the genome, we compared their sequences to the core transcript set using BLASTn, and required a perfect match over at least a 60-bp region with a transcript member of the GeneHub gene. Second, protein sequences from UniProt and PDB were compared with the GeneHub DNA sequences using the BLASTx program. For a protein record to be linked to a GeneHub gene, we required >98% identity over an at least 35 amino-acid long region. The thresholds of BLAST cutoffs were empirically determined so that false linking was minimized without over-sacrificing true signal. Third, we added links that were based on existing annotations. For example, the gene2refseq file downloaded from NCBI contains relationships between Entrez Gene records and RefSeq and GenBank accessions, so an Entrez Gene record could be linked to a GeneHub gene if a RefSeq or GenBank accession was already part of the GeneHub gene.

Gene annotation and ortholog linking

Once we built the GeneHub gene collection associated with gene and protein identifiers from various databases, it became straightforward to collect and integrate gene annotation information from multiple databases. Useful information such as gene description, accession, name and synonyms were extracted for each GeneHub gene and stored in a common database field for text searching and gene characterization purposes (Supplementary Figure 1).

The ortholog linkings between human and mouse GeneHub genes were based on the hmlg_ftp.txt file from HomoloGene (www.ncbi.nlm.nih.gov/HomoloGene/) Release 50.1. We used the orthologous Entrez Gene pairs of human and mouse if they were established by reciprocal best match between three or more organisms, or reciprocal best match, or sequence similarity with match identity >70%. We were able to link 15 868 human GeneHub genes to their mouse counterpart.

EST data collection and cleansing

EST data were downloaded from NCBI (http://www.ncbi.nlm.nih.gov/dbEST/) and processed following a previous protocol to retain only non-normalized, non-subtracted usable cancer and normal libraries (7). From the 11 July 2006 release of dbEST, we collected 4 175 880 human ESTs from all usable libraries, including 1 912 573 ESTs in 1 995 normal libraries and 2 263 307 ESTs in 3 812 cancer libraries. This represents a 35% increase in total usable EST counts from 2 years ago. For mouse, we collected 1 879 315 total ESTs, including 1 461 621 from 285 normal libraries and 417 694 from 37 cancer libraries. See Supplementary Table 2 for total number of ESTs in each category and tissue type.

EST-based expression analysis

We first mapped EST sequences to their genomic locations using GMAP (13), followed by a secondary filtering step as described earlier (>90% identity and >90% coverage). To avoid ambiguity, we discarded the ESTs (1.3% of total) that were mapped to multiple genomic loci with identical percent identity and coverage. We also eliminated those ESTs (5.5% of total) with near identical matches (with <2% difference in percent identity or coverage) to multiple genomic loci. The genomic coordinates were compared with the gene structure information (intron/exon boundary) of GeneHub genes. An EST was considered to be the product of a gene if the two entities were mapped to the same locus and share at least one exon (with minimum overlap of 30 bp). Multiple EST reads from the same clone were reduced to a single read if clone information was available. The digital expression unit (DEU) for a given gene in each tissue category is defined as the number of matching EST clones from a normalized library size of 1 million. The DEU values were calculated iteratively for each tissue type to profile expression levels across all tissues. The *Z*-test was applied to determine whether DEUs in two samples were statistically different using a method described previously (7). Comparisons were made between normal and cancer samples from the same tissue, and across different types of tissues. To improve the efficiency of the GeneHub-GEPIS program, the expression result for each of GeneHub genes was pre-computed for fast access.

For ~1000 randomly selected human genes, we compared how ESTs were mapped to genes by BLAST (≥98% identity over >60 bp region) or by the above method. For 74% of tested genes, BLAST alone would identify at least one incorrect EST, as judged by its genomic location. By design, the new method disallows this type of erroneous mapping. It is worth noting that the median DEU level for human genes in normal tissues is 50.0 based on our new method, as compared to a median level of 70.0 derived from BLAST-based EST-mapping approach (7). The average number of ESTs mapped to a gene also reduced from 128.7 to 107.5 (or a 16% reduction). The significant reduction of ESTs mapped to genes reflects the high accuracy of EST mapping by GMAP (13) and the rejection of ESTs with promiscuous matches. Manual review of EST mapping results for randomly sampled genes also confirmed the much improved mapping quality. As a confirmation step, we compared our results with a set of EST-gene mapping data independently generated by the UCSC Genome Browser team, and we found a concordance of >97%. The UCSC data are based on BLAT (16) and a series of filtering steps, and are used for the EST alignment track in the UCSC genome browser.

MicroRNA expression analysis is based on the observation that miRNA precursor sequences can be found among ESTs (17,18), and that pre-miRNA expression levels correlate with mature miRNA expression levels (19). Given this relationship, we could use EST data to approximate miRNA expression levels in various tissues. We first collected genomic locations of the miRNA stem-loop sequences from version 9.0 of miRBase (20), and then obtained all EST sequences that had any overlap with the miRNA stem-loops for expression analysis as described earlier.

The Regional GEPIS Atlas, which depicts the expression level of all genes in selected tissues for a given genomic window, was created in the similar fashion as described previously (7) but with the exception that we stored the genomic coordinates for all genes in a MySQL database instead of in plain text files.

Application implementation

The web front-end was written in HTML, javascript, CSS and Perl CGI. The Perl template module (HTML::Template) was used to achieve consistent look across different web pages. The Ajax technology was used to make the application more interactive so that text searches could be performed without leaving the input web page. We used the Prototype Javascript library (http://www.prototypejs.org/) to implement AJAX calls. This library supports AJAX interactions and provides utility functions for accessing page components and DOM manipulations. A MySQL database was used for data storage and retrieval (Supplementary Figure 1). For text-based searching, the query string can be a record identifier (e.g. accession) or gene name. The program queries DBXREF, GENE and GENE_SYNONYMS tables in sequential order to find a best match from the selected target species for the given query, regardless of the species of input record. The text search is case-insensitive and a begin-search is automatically performed if no exact match is found. All of the source code is available upon request.

PROGRAM DESCRIPTION

GeneHub-GEPIS is a tool for inferring human and mouse gene expression patterns based on normalized EST abundance in various normal and cancerous tissues. The design of the system is depicted in Figure 1. The application is composed of two parts: a front-end web interface for user input, data retrieval, display and download, and a backend engine to perform GeneHub-GEPIS analysis and data storage. The backend expression analysis relies on an integrated gene database we constructed that stores gene definitions and cross-references. Much of the documentation of this application is provided in the form of FAQs so that answers to commonly asked questions are provided on the spot.

Data input

GeneHub-GEPIS supports both text- and cDNA sequence-based data retrieval (Figure 2). For text, the application allows diverse input values ranging from gene symbols to various accession identifiers. It currently supports identifiers from common gene databases (GenBank, RefSeq, Ensembl, FANTOM, Entrez Gene, UniGene, miRBase), protein databases (PDB, UniProt) and commercial microarray platforms (Affymetrix and Agilent). For sequence-based retrieval users can either upload a single-sequence FASTA file to the web server or paste a nucleotide sequence in a text box. Users can also search for either human or mouse genes regardless the origin species of the input text and sequences. If the target species is different from the species of the input value, an ortholog search is automatically performed. The application also allows users to limit their search to a specified chromosome to avoid possible multiple matches to the input text.

It is worth noting that GeneHub-GEPIS allow direct URL access from any web pages, with a gene symbol or accession as argument. For example, when retrieving mouse *c-Met* results, the URL is: http://www.cgl.ucsf.edu/cgi-bin/genentech/genehub-gepis/web_search.pl?intype=1&xrefid=cMet&species=mouse. To obtain results for RefSeq identifier NM_001260 (human *ERBB2*), the URL is: http://www.cgl.ucsf.edu/cgi-bin/genentech/genehub-gepis/web_search.pl?intype=1&xrefid=NM_001260&species=human. Using this feature, some of the gene-based web servers, such as the widely accessed UCSC Known Genes web pages, have created links to GeneHub-GEPIS results. This is important as it can greatly increase the use of this server.

Program output

Once a unique gene match is found as one of the predefined GeneHub genes, the program directly retrieves the pre-computed EST-based gene expression results. The initial result page provides navigation links to download the result, display EST hits by libraries, and view additional graphic charts. Expression data is displayed in both a tabular format and a graphic chart (Figure 3A). The program also retrieves and displays information about input genes such as gene names, synonyms, description, genomic locations and provides links to the UCSC genome browser and others web resources such as GEO microarray results (Figure 3B). Since it is often desirable to examine the expression pattern of an orthologous gene, we provide links that allow user to quickly navigate between the result pages of human and mouse ortholog pairs. In addition, users can specify a genomic region and tissues of interest to get a Regional GEPIS Atlas view, which displays the expression values of all of the neighbor genes (Supplementary Figure 2). We also implemented a frequently requested feature that displays library information and the number of EST hits

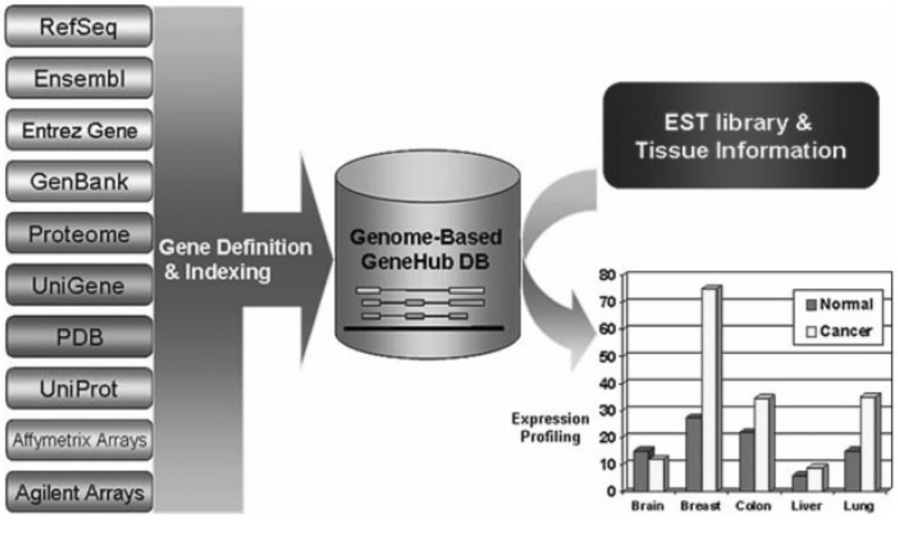


Figure 1. Conceptual representation of the design of the GeneHub-GEPIS application.

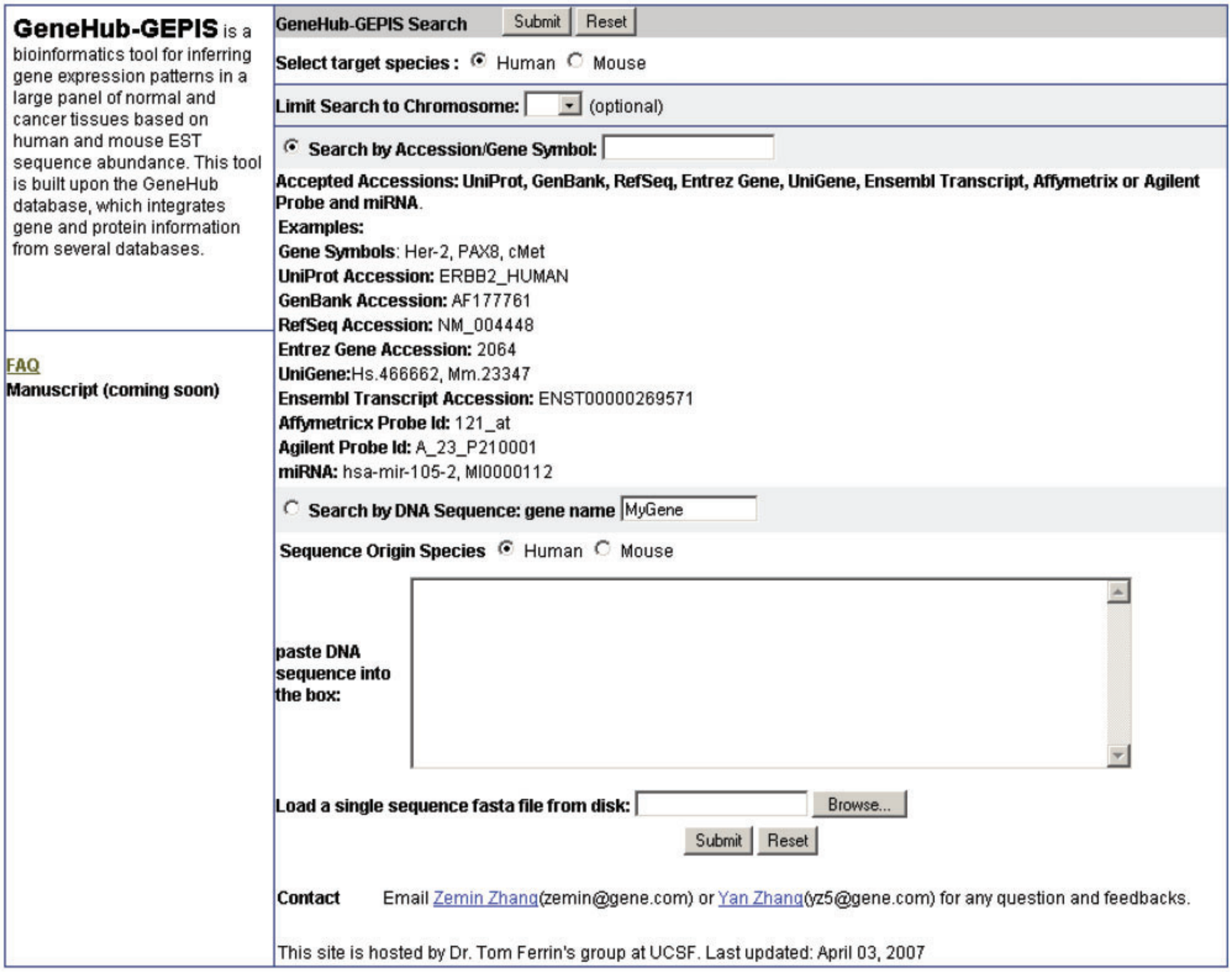
in each library (Supplementary Figure 3). The libraries are grouped by their tissue type. Due to space considerations, only libraries with matched ESTs are displayed on the web, but the full list is available for download. The EST detail page can also be bookmarked.

For text-based searching, when multiple gene matches are found, the server displays the list of all matched genes with basic information, such as gene names and genomic locations, so that users can click one of the genes for further analysis.

For sequence-based searching, the program first tries to match the input sequence with our pre-defined gene collection of the same species using BLAST (>60 bp match with >98% identity, and the top hit chosen). If a match is found in the same species, the program directly reports the matched gene's expression data. If the matched sequence is from a species different from that of the input sequence, the program queries a backend table to identify its orthologous gene for subsequent result display. In rare cases where the input sequence fails to match any pre-defined genes, the program resorts to a secondary BLAST search against the EST sequence database directly, and matched ESTs are used to compute expression results on the fly. In this case, we used the method described previously (7), and no gene annotation will be available for display.

Batch analysis and data download

To facilitate large-scale analysis, we implemented a batch analysis function and allow download of all backend data. For batch analysis, a text field is provided for pasting in a list of gene identifiers, which can be a mix of gene symbols and different types of accession numbers. For example, the input can be a list of Affymetrix microarray probe IDs supplemented with additional gene names. Upon submission, the program produces paginated result pages showing expression results for each of the input gene in a concatenated tabular format. Links to detailed information and graphical displays are provided for each gene. These tabular data, along with detailed breakdown of EST library information, can be downloaded in text files for further study.



GeneHub-GEPIS is a bioinformatics tool for inferring gene expression patterns in a large panel of normal and cancer tissues based on human and mouse EST sequence abundance. This tool is built upon the GeneHub database, which integrates gene and protein information from several databases.

FAQ
Manuscript (coming soon)

GeneHub-GEPIS Search Submit Reset

Select target species : ● Human ○ Mouse

Limit Search to Chromosome: ▾ (optional)

● **Search by Accession/Gene Symbol:**

Accepted Accessions: UniProt, GenBank, RefSeq, Entrez Gene, UniGene, Ensembl Transcript, Affymetrix or Agilent Probe and miRNA.

Examples:
Gene Symbols: Her-2, PAX8, cMet
UniProt Accession: ERBB2_HUMAN
GenBank Accession: AF177761
RefSeq Accession: NM_004448
Entrez Gene Accession: 2064
UniGene: Hs.466662, Mm.23347
Ensembl Transcript Accession: ENST00000269571
Affymetricx Probe Id: 121_at
Agilent Probe Id: A_23_P210001
miRNA: hsa-mir-105-2, MI0000112

○ **Search by DNA Sequence: gene name** MyGene

Sequence Origin Species ● Human ○ Mouse

paste DNA sequence into the box:

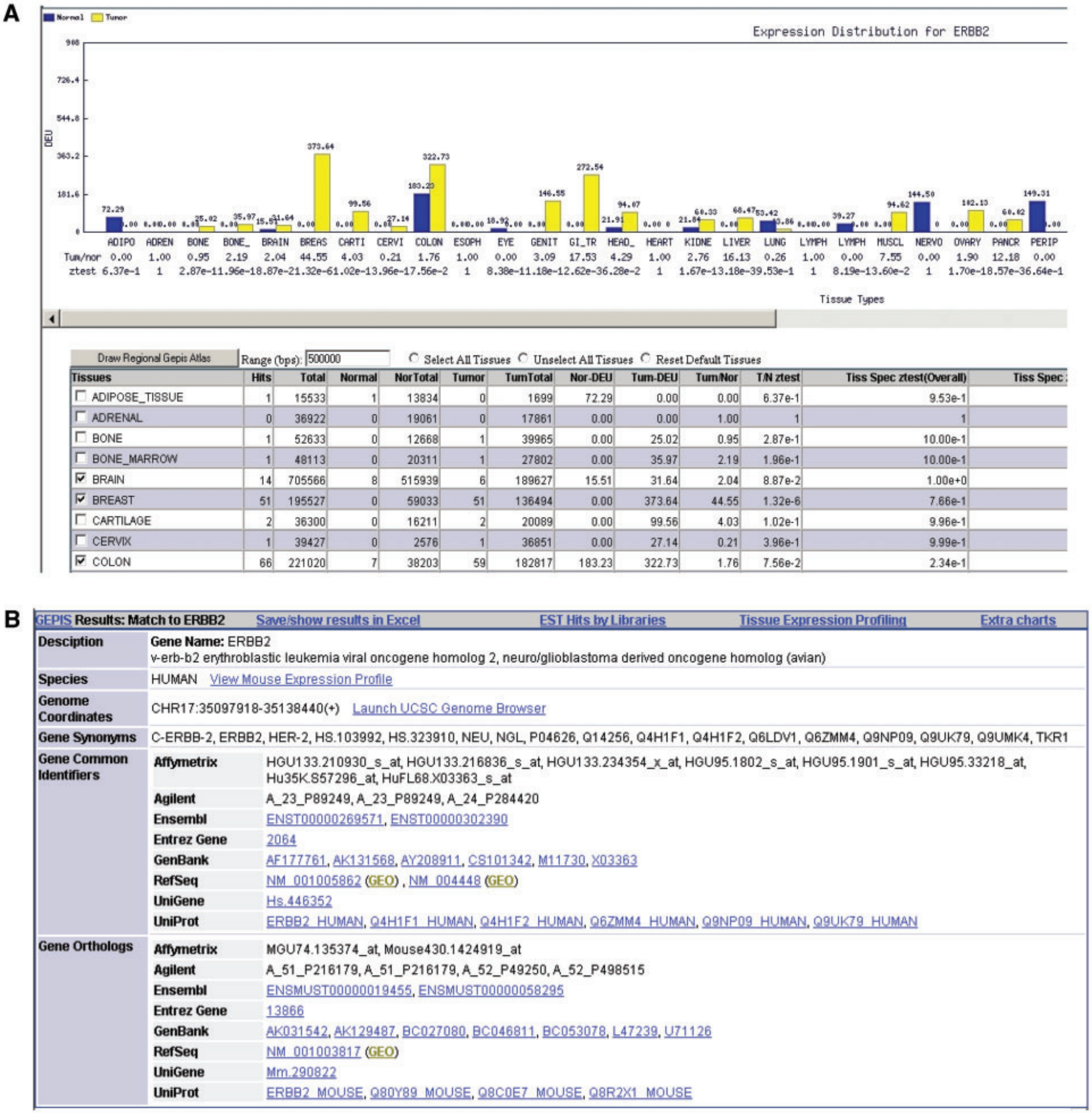
Load a single sequence fasta file from disk: Browse...

Submit Reset

Contact Email Zemin Zhang(zemin@gene.com) or Yan Zhang(yz5@gene.com) for any question and feedbacks.

This site is hosted by Dr. Tom Ferrin's group at UCSF. Last updated: April 03, 2007

Figure 2. Screenshot of GeneHub-GEPIS web input interface. Users can perform either text- or sequence-based search for a targeted organism, with an option to limit the search to a specified chromosome.


A
Normal
Tumor
Expression Distribution for ERBB2
908
726.4
544.8
363.2
181.6
0
DEU
72.29
373.64
322.73
183.23
99.56
146.55
272.54
94.07
144.50
102.13
149.31
60.02
94.62
ADIPO ADREN BONE BONE_ BRAIN BREAS CARTI CERVI COLON ESOPH EYE GENIT GI_TR HEAD_ HEART KIDNE LIVER LUNG LYMPH LYMPH MUSCL NERVO OVARY PANCR PERIP
Tum/nor 0.00 1.00 0.95 2.19 2.04 44.55 4.03 0.21 1.76 1.00 0.00 3.09 17.53 4.29 1.00 2.76 16.13 0.26 1.00 0.00 7.55 0.00 1.90 12.18 0.00
ztest 6.37e-1 1 2.87e-11.96e-18.87e-21.32e-61.02e-13.96e-17.56e-2 1 8.38e-11.18e-12.62e-36.28e-2 1 1.67e-13.18e-39.53e-1 1 8.19e-13.60e-2 1 1.70e-18.57e-36.64e-1
Tissue Types
Draw Regional Gepis Atlas
Range (bps): 500000
○ Select All Tissues ○ Unselect All Tissues ○ Reset Default Tissues
B

Tissues	Hits	Total	Normal	NorTotal	Tumor	TumTotal	Nor-DEU	Tum-DEU	Tum/Nor	T/N ztest	Tiss Spec ztest(Overall)	Tiss Spec :
□ ADIPOSE_TISSUE	1	15533	1	13834	0	1699	72.29	0.00	0.00	6.37e-1	9.53e-1	
□ ADRENAL	0	36922	0	19061	0	17861	0.00	0.00	1.00	1	1	
□ BONE	1	52633	0	12668	1	39965	0.00	25.02	0.95	2.87e-1	10.00e-1	
□ BONE_MARROW	1	48113	0	20311	1	27802	0.00	35.97	2.19	1.96e-1	10.00e-1	
☑ BRAIN	14	705566	8	515939	6	189627	15.51	31.64	2.04	8.87e-2	1.00e+0	
☑ BREAST	51	195527	0	59033	51	136494	0.00	373.64	44.55	1.32e-6	7.66e-1	
□ CARTILAGE	2	36300	0	16211	2	20089	0.00	99.56	4.03	1.02e-1	9.96e-1	
□ CERVIX	1	39427	0	2576	1	36851	0.00	27.14	0.21	3.96e-1	9.99e-1	
☑ COLON	66	221020	7	38203	59	182817	183.23	322.73	1.76	7.56e-2	2.34e-1	

GEPIS Results: Match to ERBB2 Save/show results in Excel EST Hits by Libraries Tissue Expression Profiling Extra charts

Description	Gene Name: ERBB2 v-erb-b2 erythroblastic leukemia viral oncogene homolog 2, neuro/glioblastoma derived oncogene homolog (avian)	
Species	HUMAN View Mouse Expression Profile	
Genome Coordinates	CHR17:35097918-35138440(+) Launch UCSC Genome Browser	
Gene Synonyms	C-ERBB-2, ERBB2, HER-2, HS.103992, HS.323910, NEU, NGL, P04626, Q14256, Q4H1F1, Q4H1F2, Q6LDV1, Q6ZMM4, Q9NP09, Q9UK79, Q9UMK4, TKR1	
Gene Common Identifiers	Affymetrix	HGU133.210930_s_at, HGU133.216836_s_at, HGU133.234354_x_at, HGU95.1802_s_at, HGU95.1901_s_at, HGU95.33218_at, Hu35K.S57296_at, HuFL68.X03363_s_at
	Agilent	A_23_P89249, A_23_P89249, A_24_P284420
	Ensembl	ENST00000269571, ENST00000302390
	Entrez Gene	2064
	GenBank	AF177761, AK131568, AY208911, CS101342, M11730, X03363
	RefSeq	NM_001005862 (GEO) , NM_004448 (GEO)
	UniGene	Hs.446352
	UniProt	ERBB2_HUMAN, Q4H1F1_HUMAN, Q4H1F2_HUMAN, Q6ZMM4_HUMAN, Q9NP09_HUMAN, Q9UK79_HUMAN
Gene Orthologs	Affymetrix	MGU74.135374_at, Mouse430.1424919_at
	Agilent	A_51_P216179, A_51_P216179, A_52_P49250, A_52_P498515
	Ensembl	ENSMUST00000019455, ENSMUST00000058295
	Entrez Gene	13866
	GenBank	AK031542, AK129487, BC027080, BC046811, BC053078, L47239, U71126
	RefSeq	NM_001003817 (GEO)
	UniGene	Mm.290822
	UniProt	ERBB2_MOUSE, Q80Y89_MOUSE, Q8C0E7_MOUSE, Q8R2X1_MOUSE

Figure 3. Screenshots of GeneHub-GEPIS web output pages. (**A**). Tissue Expression Profile Chart and GeneHub-GEPIS result table. The bar chart displays the normal (blue) and tumor (yellow) DEU values of each type of tissue, and the table shows numeric data and statistics. The user can select tissues and specify a genomic range in this page to draw Regional GEPIS Atlas. (**B**) The gene summary section provides a short description of gene function, species, genomic coordinates and synonyms. It provides a link to navigate to the result for a gene's ortholog, links to download results, view EST hits by libraries, and view additional graphic charts, and links to additional web resources such as the UCSC genome browser.

Since the backend data can be potentially useful for other purposes, we provide a download page where all backend data can be retrieved. These include gene mapping and cross-reference data, exon and boundary definitions, EST mapping and associated library information, pre-computed expression results for all pre-defined genes, and detailed EST library distribution information for each gene. Such files can be used by power users for global surveys of expression across a large number of genes.

CONCLUDING REMARKS

Despite the increasing prevalence of microarray data, ESTs remain as a significant source of data for expression analysis, and can provide benefits over microarrays in some cases (7). However, the value of EST data can only be fully realized with the availability of powerful and user-friendly tools that transform loosely organized EST information into meaningful expression results. Guided by input from the user community, we aimed to make

GeneHub-GEPIS reliable, powerful, easy to use and widely accessible. At this point, GeneHub-GEPIS can report estimated expression levels in about 40 different types of normal and cancerous tissues for a given gene or a list of genes. As more EST data become available, it will be possible to analyze gene expression in more detailed tissue subtypes and for additional organisms. We have noticed a dramatic increase of traffic to our web server since the release of GeneHub-GEPIS, and we hope that GeneHub-GEPIS will stimulate greater usage of EST data and perhaps additional software development in this area.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank the UCSF Computer Graphics Laboratory and Dr Thomas Ferrin for hosting this web server. We thank Fan Hsu and Jim Kent from UCSC for early evaluation and comments. We thank our Genentech colleagues Thomas Wu and Colin Watanabe for early access to the GMAP program, Josh Kaminker, Pete Haverty, Jerry Zhijun Tang, Hilary Clark, Reece Hart, Kiran Mukhyala and Jim Fitzgerald for helpful discussions and computational support. Funding to pay the Open Access publication charges for this article was provided by Genentech, Inc.

Conflict of interest statement. None declared.

REFERENCES

1. Adams,M.D., Kerlavage,A.R., Fields,C. and Venter,J.C. (1993) 3,400 new expressed sequence tags identify diversity of transcripts in human brain. *Nat. Gen.*, **4**, 256–267.
2. Hishiki,T., Kawamoto,S., Morishita,S. and Okubo,K. (2000) BodyMap: a human and mouse gene expression database. *Nucleic Acids Res.*, **28**, 136–138.
3. Scheurle,D., DeYoung,M.P., Binninger,D.M., Page,H., Jahanzeb,M. and Narayanan,R. (2000) Cancer gene discovery using digital differential display. *Cancer Res.*, **60**, 4037–4043.
4. Brentani,H., Caballero,O.L., Camargo,A.A., da Silva,A.M., da Silva,W.A.Jr., Dias Neto,E., Grivet,M., Gruber,A., Moreira Guimaraes,P.E. *et al.* (2003) The generation and utilization of a cancer-oriented representation of the human transcriptome by using expressed sequence tags. *Proc. Natl Acad. Sci. USA*, **100**, 13418–13423.
5. Wu,X., Walker,M.G., Luo,J. and Wei,L. (2005) GBA server: EST-based digital gene expression profiling. *Nucleic Acids Res.*, **33**, W673–676.
6. Ferguson,D.A., Chiang,J.T., Richardson,J.A. and Graff,J. (2005) eXPRESSION: an in silico tool to predict patterns of gene expression. *Gene Expr. Patterns*, **5**, 619–628.
7. Zhang,Y., Eberhard,D.A., Frantz,G.D., Dowd,P., Wu,T.D., Zhou,Y., Watanabe,C., Luoh,S.M., Polakis,P. *et al.* (2004) GEPIS–quantitative gene expression profiling in normal and cancer tissues. *Bioinformatics*, **20**, 2390–2398.
8. D'Agostino,N., Aversano,M., Frusciante,L. and Chiusano,M.L. (2007) TomatEST database: in silico exploitation of EST data to explore expression patterns in tomato species. *Nucleic Acids Res.*, **35**, D901–D905.
9. Uenishi,H., Eguchi-Ogawa,T., Shinkai,H., Okumura,N., Suzuki,K., Toki,D., Hamasima,N. and Awata,T. (2007) PEDE (Pig EST Data Explorer) has been expanded into Pig Expression Data Explorer, including 10 147 porcine full-length cDNA sequences. *Nucleic Acids Res.*, **35**, D650–D653.
10. Kim,H., Lim,D., Han,B.K., Sung,S., Jeon,M., Moon,S., Kang,Y., Nam,J. and Han,J.Y. (2006) ChickGCE: a novel germ cell EST database for studying the early developmental stage in chickens. *Genomics*, **88**, 252–257.
11. Houde,M., Belcaid,M., Ouellet,F., Danyluk,J., Monroy,A.F., Dryanova,A., Gulick,P., Bergeron,A., Laroche,A. *et al.* (2006) Wheat EST resources for functional genomics of abiotic stress. *BMC Genomics*, **7**, 149.
12. Chini,V., Rimoldi,S., Terova,G., Saroglia,M., Rossi,F., Bernardini,G. and Gornati,R. (2006) EST-based identification of genes expressed in the liver of adult seabass (*Dicentrarchus labrax, L.*). *Gene*, **376**, 102–106.
13. Wu,T.D. and Watanabe,C.K. (2005) GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics*, **21**, 1859–1875.
14. Zhang,Y., Li,J., Kong,L., Gao,G., Liu,Q.R. and Wei,L. (2007) NATsDB: Natural Antisense Transcripts DataBase. *Nucleic Acids Res.*, **35**, D156–D161.
15. Engstrom,P.G., Suzuki,H., Ninomiya,N., Akalin,A., Sessa,L., Lavorgna,G., Brozzi,A., Luzi,L., Tan,S.L. *et al.* (2006) Complex loci in human and mouse genomes. *PLoS Gen.*, **2**, e47.
16. Kent,W.J. (2002) BLAT–the BLAST-like alignment tool. *Genome Res.*, **12**, 656–664.
17. Smalheiser,N.R. (2003) EST analyses predict the existence of a population of chimeric microRNA precursor-mRNA transcripts expressed in normal human and mouse tissues. *Gen. Biol.*, **4**, 403.
18. Zhang,B.H., Pan,X.P., Wang,Q.L., Cobb,G.P. and Anderson,T.A. (2005) Identification and characterization of new plant microRNAs using EST analysis. *Cell Res.*, **15**, 336–360.
19. Thomson,J.M., Newman,M., Parker,J.S., Morin-Kensicki,E.M., Wright,T. and Hammond,S.M. (2006) Extensive post-transcriptional regulation of microRNAs and its implications for cancer. *Genes Dev.*, **20**, 2202–2207.
20. Griffiths-Jones,S., Grocock,R.J., van Dongen,S., Bateman,A. and Enright,A.J. (2006) miRBase: microRNA sequences, targets and gene nomenclature. *Nucleic Acids Res.*, **34**, D140–D144.