



GEPIS—quantitative gene expression profiling in normal and cancer tissues

Yan Zhang¹, David A. Eberhard², Gretchen D. Frantz², Patrick Dowd⁴, Thomas D. Wu¹, Yan Zhou¹, Colin Watanabe¹, Shih-Ming Luoh¹, Paul Polakis³, Kenneth J. Hillan², William I. Wood¹ and Zemin Zhang^{1,*}

¹Department of Bioinformatics, ²Department of Pathology, ³Department of Molecular Oncology and ⁴Department of Molecular Biology, Genentech Inc., South San Francisco, CA 94080, USA

Received on January 12, 2004; revised on March 1, 2004; accepted on March 2, 2004

Advance Access publication April 8, 2004

ABSTRACT

Motivation: Expression profiling in diverse tissues is fundamental to understanding gene function as well as therapeutic target identification. The vast collection of expressed sequence tags (ESTs) and the associated tissue source information provides an attractive opportunity for studying gene expression.

Results: To facilitate EST-based expression analysis, we developed GEPIS (gene expression profiling *in silico*), a tool that integrates EST and tissue source information to compute gene expression patterns in a large panel of normal and tumor samples. We found EST-based expression patterns to be consistent with published papers as well as our own experimental results. We also built a GEPIS Regional Atlas that depicts expression characteristics of all genes in a selected genomic region. This program can be adapted for large-scale screening for genes with desirable expression patterns, as illustrated by our large-scale mining for tissue- and tumor-specific genes.

Availability: The email server version of the GEPIS application is freely available at <http://share.gene.com/share/gepis>. An interactive version of GEPIS will soon be freely available at <http://www.cgl.ucsf.edu/Research/genentech/gepis/>. The source code, modules, data and gene lists can be downloaded at <http://share.gene.com/share/gepis>

Contact: zemin@gene.com

Supplementary information: Supplementary tables and figures are available at <http://www.cgl.ucsf.edu/Research/genentech/gepis/>

INTRODUCTION

Expression analysis is essential for understanding gene functions and identifying therapeutic targets. A gene with a restricted expression pattern is likely to be functionally related

to the tissue or disease state in which this gene is specifically expressed. The development of high-throughput expression analysis technologies such as microarrays has dramatically improved our ability to study the expression of a large panel of genes in a particular tissue. Meanwhile, it is equally desirable to study the expression of a single gene over a wide spectrum of tissue types. For example, comprehensive expression profiling helps identify targets for monoclonal antibody-based cancer therapies (Green *et al.*, 2000), as nonspecific expression of the target gene would predict undesirable toxicity. Although it is possible to perform multiplexed cDNA- or oligonucleotide-based microarray experiments over many different tissues to derive gene expression profiles (Ramaswamy *et al.*, 2001; Ross *et al.*, 2000; Su *et al.*, 2002), it is important to develop methods that profile expression of all genes, not just genes on chips, in a simple and quantitative manner.

The collection of expressed sequence tags (ESTs) (Adams *et al.*, 1991) provides an increasingly attractive source for expression analysis. Since EST clone frequency is in principle proportional to the corresponding gene's expression level (Adams *et al.*, 1993), ESTs have been successfully used previously for studying expression signatures or differential expression analysis (Audic and Claverie, 1997; Ewing *et al.*, 1999; Hishiki *et al.*, 2000; Scheurle *et al.*, 2000; Schmitt *et al.*, 1999). More recently, the application of ESTs in studying the human cancer transcriptome has been summarized by reported large-scale efforts by the Cancer Genome Anatomy Project (CGAP) and Human Cancer Genome Project (HCGP) (Brentani *et al.*, 2003). However, unlike microarray technology, the EST-based method has yet to become a prevalent approach for expression analysis, as it has been limited by several factors: insufficient EST data for representing diverse tissues, concerns over subtracted and normalized libraries, the need for experimental validation of EST-derived results and a dearth of user-friendly tools for analyzing expression

*To whom correspondence should be addressed.

results. With increasingly comprehensive EST collections and better-annotated EST libraries, it is now possible to develop EST-based technology that performs thorough and reliable gene expression analysis. In particular, EST data representing both normal tissues and their cancer counterparts would facilitate both tissue- and tumor-specificity analysis of essentially all genes.

In this paper, we describe GEPIS (gene expression profiling *in silico*), a publicly available web application, which computes expression profiles of input sequences in normal and cancer tissues. In addition, to facilitate expression analysis in the genomic context, we provide a graphical expression atlas of a genomic region that shows the expression patterns of the adjacent genes, as the expression of neighboring genes often sheds light into the mechanisms of differential gene expression (Spellman and Rubin, 2002; Zhou *et al.*, 2003). Furthermore, we provide the backend GEPIS program that can be adapted for large-scale data mining. To illustrate the capability of GEPIS-based mining, we screened for tissue-specific as well as tumor-specific genes, and experimentally validated targets found by GEPIS. This microarray-independent method will be a valuable resource for studying gene expression patterns, for screening for targets with desirable specificities, and for gaining insights into mechanisms of differential gene expression.

METHODS

Collection and classification of EST libraries

The human EST collection consists of all sequences in the EST division of GenBank (<http://www.ncbi.nlm.nih.gov/dbEST/>) (Boguski *et al.*, 1993). Quality information contained in the GenBank file was used to trim low-quality bases from sequences. Based on tissue and histology data for library information from the National Cancer Institute (NCI) and CGAP (<http://cgap.nci.nih.gov/Tissues/> and ftp://ftp1.nci.nih.gov/pub/CGAP/Hs_LibData.dat), each EST was assigned a tissue source value and a disease value.

Since the classification and selection of EST libraries were expected to be critical for reliable expression analysis, we established a number of quality control steps to identify usable libraries and to group them in the appropriate tissue categories. First, we removed ESTs with unknown, ambiguous or pooled tissue sources and discarded ESTs that failed to fall into either the 'normal' or the 'cancer' categories. For metastatic tumors, the primary tissue of origin was recorded as the tumor source. ESTs from libraries labeled as 'normalized' or 'subtracted' were also excluded as they might obscure EST abundance calculation. In addition, 118 libraries derived from embryonic or fetal tissues were removed from this study as they might exhibit different expression signatures from their adult counterparts. Finally, we discarded a number of libraries that might have been misclassified according to our expression analyses. For instance, library NCI_CGAP_Br16, annotated as a mammary gland library, was atypical of other breast

tissue libraries because it contained a large number of ESTs for prostate-specific genes such as prostate-specific antigen (PSA) and acid phosphatase, prostate (ACPP). This discrepancy was most likely caused by a rare problem in library annotation, but it alerted us to use the annotation information with caution. These data cleansing steps led to a lookup table that contains linking information from ESTs to libraries to tissue types.

Gene expression profiling *in silico*

A computational program, GEPIS, was developed to calculate gene expression level and tissue distribution based on EST data and library information. First, the sequence alignment program BLAST (Altschul *et al.*, 1990) was used to find EST sequences that matched a human gene of interest. Multiple EST reads from the same clone were reduced to a single read. The library identifier of each EST clone was then used to associate its tissue source based on the library-tissue lookup table. Multiple libraries of the same tissue type were aggregated in both the number of gene-matched EST clones and total number of EST clones. This allowed a digital expression unit (DEU) value to be derived for each tissue category, calculated as the total number of matching EST clones divided by the sum of library sizes and multiplied by 1 000 000. Even though a typical EST library contained only a few thousand-clone sequences, multiple libraries from the same tissue type were aggregated to form a composite data source that became more information-rich and better represented rare genes. DEU calculation thus approximates EST representation in libraries of a pool of one million sequences. DEU values were calculated iteratively for each tissue type to profile expression levels across all tissues.

The *Z*-test was applied to determine whether DEU in sample type A was statistically higher than DEU in sample type B. Comparisons could be made between normal and cancer samples from the same tissue, or between two different types of tissues. For a given gene, the common relative abundance \hat{p} was computed in all libraries by taking the sum of clone counts for the gene over all libraries and dividing by the total number of clones over all libraries. The relative abundance for the gene in type A (\hat{p}_A) and in type B (\hat{p}_B) libraries was also calculated. The test statistic *Z*-score was calculated as follows:

$$Z = \frac{\hat{p}_A - \hat{p}_B}{\sqrt{\hat{p}(1 - \hat{p})[(1/N_A) + (1/N_B)]}}$$

where N_A and N_B are the total numbers of EST clones derived from A and B. The *Z*-score can then be referred to the normal distribution to yield a *P*-value.

All source code is available at <http://www.gene.com/share/gepis> along with information on how the program can be set-up. The backend program runs in the UNIX environment and we provide CGI scripts for setting up a web interface that displays results graphically. The response time of the

program is primarily dependent upon the speed of the BLAST program against the EST database.

Tissue-specificity searching and display

To search for genes with specific tissue expression, we performed a series of inter-tissue *Z*-tests of their normal DEU units for all human genes available in the LocusLink (Pruitt and Maglott, 2001) database. The 23 995 LocusLink entries for human genes (downloaded from ftp://ftp.ncbi.nih.gov/refseq/LocusLink/LL_tmpl.gz, November 2003) with representative mRNA sequences were included in our analyses. The genomic coordinates of these genes (Build-34 of the human genome) were retrieved from NCBI (ftp://ftp.ncbi.nih.gov/refseq/H_sapiens/H_sapiens/maps/mapview/BUILD.34/seq_gene.md.gz). For each gene, we compared the tissue with the highest DEU level against each of the remaining tissues. To be considered as tissue specific in this study, a gene needed to satisfy a number of stringent selection criteria. When comparing each pair of tissues, we required a minimum of 2-fold difference in expression level and a *P*-value of <0.025 from the *Z*-test. In addition, we restricted target tissues to those with a combined library size of over 10 000 ESTs. Moreover, since the median DEU level in all normal tissues across all human LocusLink genes was ~ 70 , we required a DEU level over 140 in the target tissue and a DEU level under 70 in the remaining tissues.

Genes differentially expressed in cancer samples

Ideal therapeutic targets for tumors are often gene products that are up-regulated in tumor samples when compared with matched normal. It is also desirable that the target genes show lower expression in all other normal tissues to minimize toxicity. We performed GEPIS analysis for all genes available in the LocusLink dataset and collected genes that satisfied the following selection criteria. First, the expression level in the target tumor tissue was at least 2-fold higher than the corresponding normal tissue with a *P*-value of <0.025 from the *Z*-test. In addition, both the target tumor and normal tissues had a total library size of over 10 000 ESTs. Furthermore, guided by the statistical 'rule of threes' (Louis, 1981), we required a minimum of three EST hits from the target tumor tissue to increase the reliability of our results. Finally, the DEUs in the rest of the normal tissues had to be under 70, the approximate median expression level.

TaqMan analysis of gene expression levels

RNA was extracted from frozen samples of nine colon adenocarcinomas and two normal colon tissues and purified by CsCl gradient centrifugation, phenol/chloroform extraction and ethanol precipitation. Quantitative polymerase chain reactions (PCRs) were performed under standard conditions (Heid *et al.*, 1996). Amplification primers and hybridization probes were designed for each gene in the 3'-untranslated region of the cDNA. Data were analyzed by the Comparative Ct Method

using GAPDH as the internal control (Applied Biosystems User Bulletin # 2 Relative Quantitation of Gene Expression 1997).

Creating Regional GEPIS Atlas

Regional GEPIS Atlas is a composite of visual transcriptome maps depicting the expression level of all neighboring genes in selected tissues. To do this, we pre-computed the GEPIS results for all representative mRNA sequences for human LocusLink and stored their genomic coordinates downloaded from NCBI. For an input sequence, the web application first identifies the matched LocusLink record by BLAST analysis (minimum requirement of match length of >60 bp with $>98\%$ identity, and the top hit was chosen) and then retrieves the surrounding genes residing within a user-specified distance (in kb). The precomputed $\log_2(\text{DEU})$ values for both normal and tumor tissues of each neighboring gene are then plotted along the chromosome according to its genomic coordinates, producing local transcriptome maps in each of the user-selected tissues. The drawing program was implemented in Perl and runs on the UNIX platform, and each chart is saved in the PNG format. The drawing program required extending the Perl GD and GD::Graph (<http://www.cpan.org>) modules. The graphical tools can be downloaded at <http://www.gene.com/share/gepis>

RESULTS

EST-based profiling of gene expression in diverse tissues

The steady accumulation of ESTs in the public database has made it feasible to reliably profile gene expression in a reasonably comprehensive manner. After excluding EST libraries that were unsuitable for expression analysis, we found that usable EST libraries represented 43 diverse tissue categories, with most of the tissues covered by both normal and tumor types (Supplementary Table 1). Using GEPIS, gene expression profiles were rapidly generated by retrieving the number of matching EST clones for a given gene from each library for all available tissue types. The expression level of each gene is given as a DEU value, the number of matching EST clones per million total library clones.

Digital expression unit should be directly proportional to the absolute level of gene expression, i.e. the copy number of mRNA per cell. We examined whether our expression profiles agree with previous experimental data for well-characterized genes. In general, EST-based expression profile correlated very well with previous observations. For example, PSA is known to be specifically expressed in prostate (Clements *et al.*, 2001; Diamandis *et al.*, 2000). In EST-based expression profiles, PSA exhibited remarkable tissue specificity in prostate, with over 90% of all matching ESTs falling in this tissue category (Fig. 1A). BEHAB/brevican is another well-characterized gene shown to be specific to

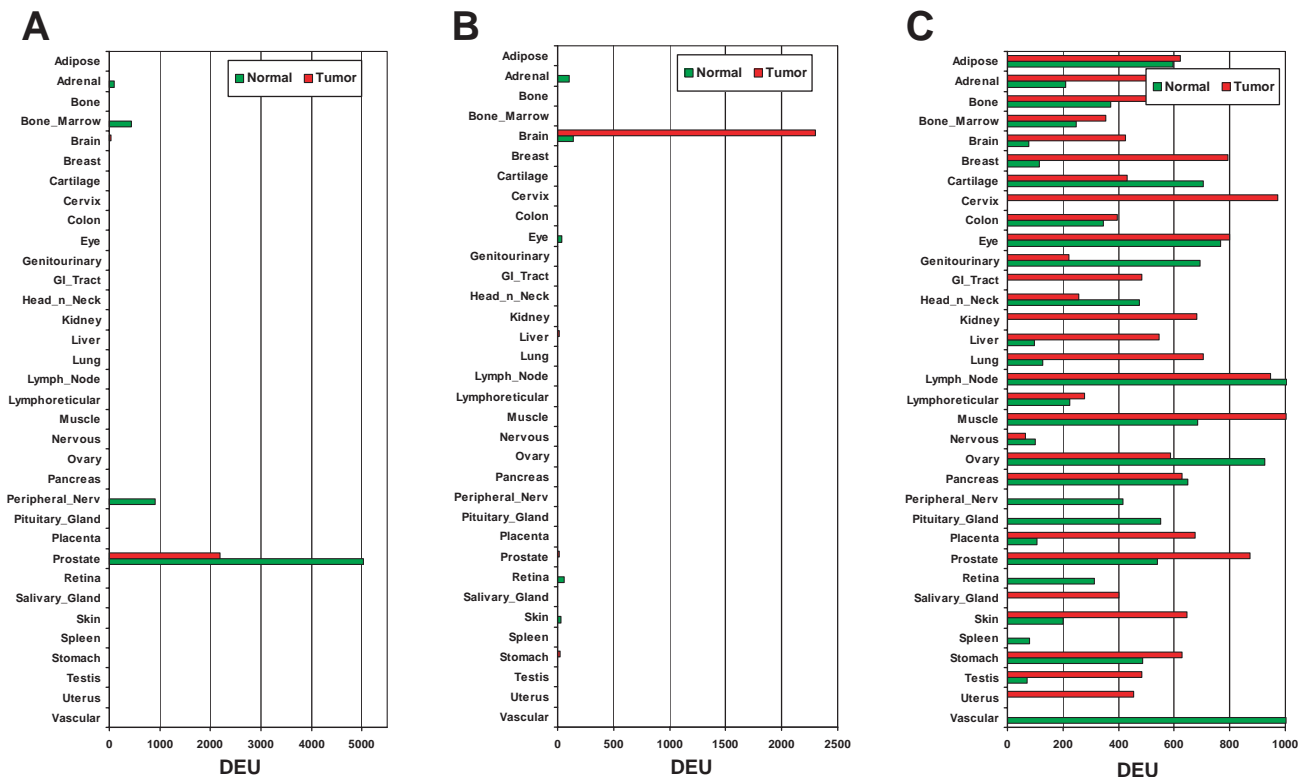


Fig. 1. EST-based expression profiles of PSA (A), BEHAB (B) and RPL12 (C). The ‘digital’ expression levels, represented by DEU units, in normal and tumor samples are shown in green and red, respectively. Although a more comprehensive profile covering additional tissues and disease types could be generated, only a standard set of 34 tissues are shown here whose combined EST library sizes are over 10 000.

brain and is up-regulated in invasive glioma (Gary *et al.*, 2000; Jaworski *et al.*, 1996; Viapiano *et al.*, 2003). GEPIS again faithfully recapitulated such specific expression pattern (Fig. 1B) for BEHAB/brevican. In contrast, other housekeeping genes, such as RPL12 (ribosomal protein L12), exhibited ubiquitous expression by GEPIS analysis in all tissues where EST data are available (Fig. 1C).

DEU correlates with TaqMan-determined expression levels

Since the GEPIS-derived expression level in a particular tissue was based on a pool of different EST libraries and many of original tissue samples were not readily available, it was impractical for us to experimentally validate whether our results directly correlated with true mRNA levels in the original tissue samples. Instead, we obtained 11 normal and malignant colon samples from a commercial source, and determined whether our GEPIS results reflect mRNA levels in these colon samples. A series of real-time PCRs (TaqMan) were performed for 40 different genes that were expected to represent high, moderate and low expression levels in colon. The TaqMan CT values of the genes tested were plotted against the log of the DEU value based on ESTs from colon libraries, as a difference of 1 CT unit corresponds to a 2-fold

difference in mRNA (Fig. 2). In spite of the intrinsic variation of TaqMan analysis and the use of different tissue samples in the TaqMan and EST analyses, the DEU values correlated well with TaqMan-determined expression levels. Over a broad range of expression levels, the CT values correlated with the \log_2 (DEU) in a linear fashion with a slope of approximately -1.0 and an R^2 of 0.6 (Fig. 2). Similar correlations were also observed when we performed TaqMan analyses using breast and prostate samples (data not shown). These analyses provide experimental evidence for the validity of EST-based expression analysis.

GEPIS screening for tissue-specific genes

The ability to examine gene expression rapidly and reliably in a wide spectrum of tissues has many useful applications. One such application is the large-scale identification of genes with specific tissue expression, as GEPIS successfully recapitulated expected expression patterns of known tissue-specific genes such as PSA and BEHAB/brevican (Fig. 1). To screen for tissue-specific genes, we evaluated all available genes computationally using criteria similar to those used in previous work (Su *et al.*, 2002). The median DEU level of all genes was ~ 70 . A gene was considered tissue-specific if its DEU level was above 140 in one normal tissue, but less than

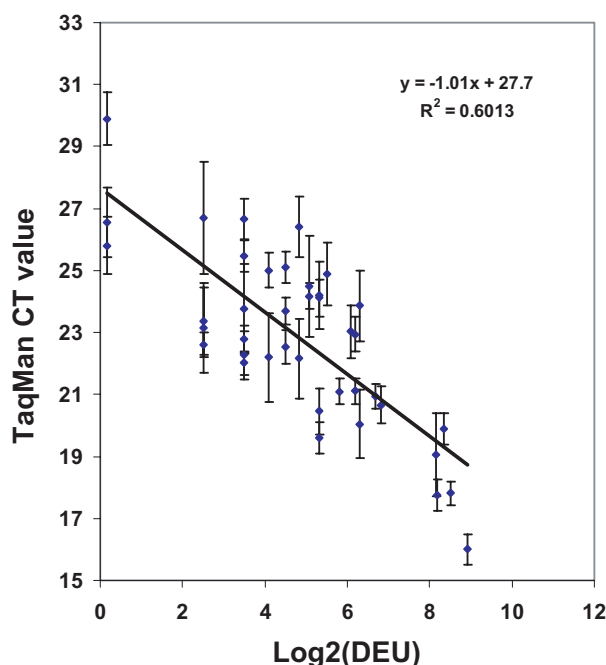


Fig. 2. Correlation between TaqMan CT values and EST-based expression levels. Each dot represents one of the 40 tested genes that were expected to exhibit wide ranges of expression levels in colon samples, such as the abundant gene GAPDH and the rare gene GNRHR. The TaqMan CT values were the averages from 11 different colon samples (two normal and nine tumor samples), and error bars show SDs. DEU values were based on normal and tumor colon EST libraries. Since an increase of 1 CT value reflects a 2-fold decrease in expression level, the averaged CT values are plotted against the base-2 logarithms of DEU. A few genes that did not have any EST representation from colon libraries were given a pseudo clone count of 0.2 to be included in the log plot. The linear trend line has slope of -1.01 and y -intercept of 27.7 . The R^2 value is 0.60 .

70 in all other normal tissues, and the difference was statistically significant using Z -tests. From a total of $\sim 24\,000$ human genes in the LocusLink collection, 248 genes passed our selection criteria (Supplementary Table 2).

The expression profiles for these 248 tissue-specific genes are shown in a heat-map representation (Supplementary Figure 1). Many of these genes are well-known to be specifically expressed in one unique tissue, such as insulin in pancreas and crystallin in eye. Based on gene annotations, the collections of tissue-specific genes are consistent with the biological functions of each tissue. For example, many of the kidney-specific genes encode for protein products that are solute carriers and ion channels and many of the 26 pancreas-specific genes produce insulin and a variety of proteases (Fig. 3). To confirm this further, we examined the pancreas-specific genes to determine whether they were consistent with previous observations. A literature survey of these 26 genes indicated that 20 genes showed pancreas-specific

expression or had pancreas-specific functions and 3 genes showed dual expression specificity in brain and pancreas. One gene had conflicting expression data in the literature, and there was no information for the remaining two uncharacterized genes. These results underscore the reliability of GEPIS-based screening for tissue-specific genes and will shed light into the biological functions of the uncharacterized genes.

Large-scale GEPIS identification of cancer target genes

The abundant representation of both normal and tumor samples in the EST database also enables large-scale searching for genes differentially expressed in tumor samples. The identification of these 'tumor genes' will not only help us to understand their molecular functions, but may also provide candidate targets for cancer therapies. To illustrate the capability of GEPIS in this regard, we screened all genes available in the LocusLink dataset and collected genes that showed 2-fold higher expression when comparing tumor to normal tissues with statistical significance measured by Z -test. For toxicity considerations, we excluded those genes showing high expression (above the median level) in any of the remaining normal tissues. This procedure produced 370 genes that exhibited differential expression in 18 different tissues (Supplementary Figure 2 and Supplementary Table 3). The expression patterns in a panel of normal and tumor samples for some of these 'tumor genes' are shown in a heat-map representation (Fig. 4). Many of the genes displayed ideal expression profile for cancer targets: specific expression in one of the tumor types. We found expected genes, such as BEHAB/brevican from brain tumors (Gary *et al.*, 2000; Jaworski *et al.*, 1996; Viapiano *et al.*, 2003), and several melanoma antigen family (MAGE) members from skin cancer samples (Basarab *et al.*, 1999; Brasseur *et al.*, 1995; Chen *et al.*, 1997; Lucas *et al.*, 1998). This list of 'cancer genes' will be a valuable resource for cancer research and target discovery.

GEPIS gene expression in the genomic context

Analysis of gene expression in a genomic context often provides additional insights into the basis for any observed differential expression pattern. In fact, the GEPIS tool has been successfully applied to the identification of potential tumor amplicons (Zhou *et al.*, 2003), as aberrant DNA amplification often results in clusters of genes along the genome that display up-regulation in cancer samples. Our GEPIS application creates a Regional GEPIS Atlas that displays the expression pattern of all genes in the vicinity of an input DNA sequence. One such example is shown in Supplementary Figure 3 where the normal/tumor expression patterns in selected tissues are shown for 16 genes located in a 1.4 Mb genomic window centered on the input *MMP1* gene. The length of the genomic window and tissues of interest can be adjusted. This display quickly illustrates not only the proximity of nine members

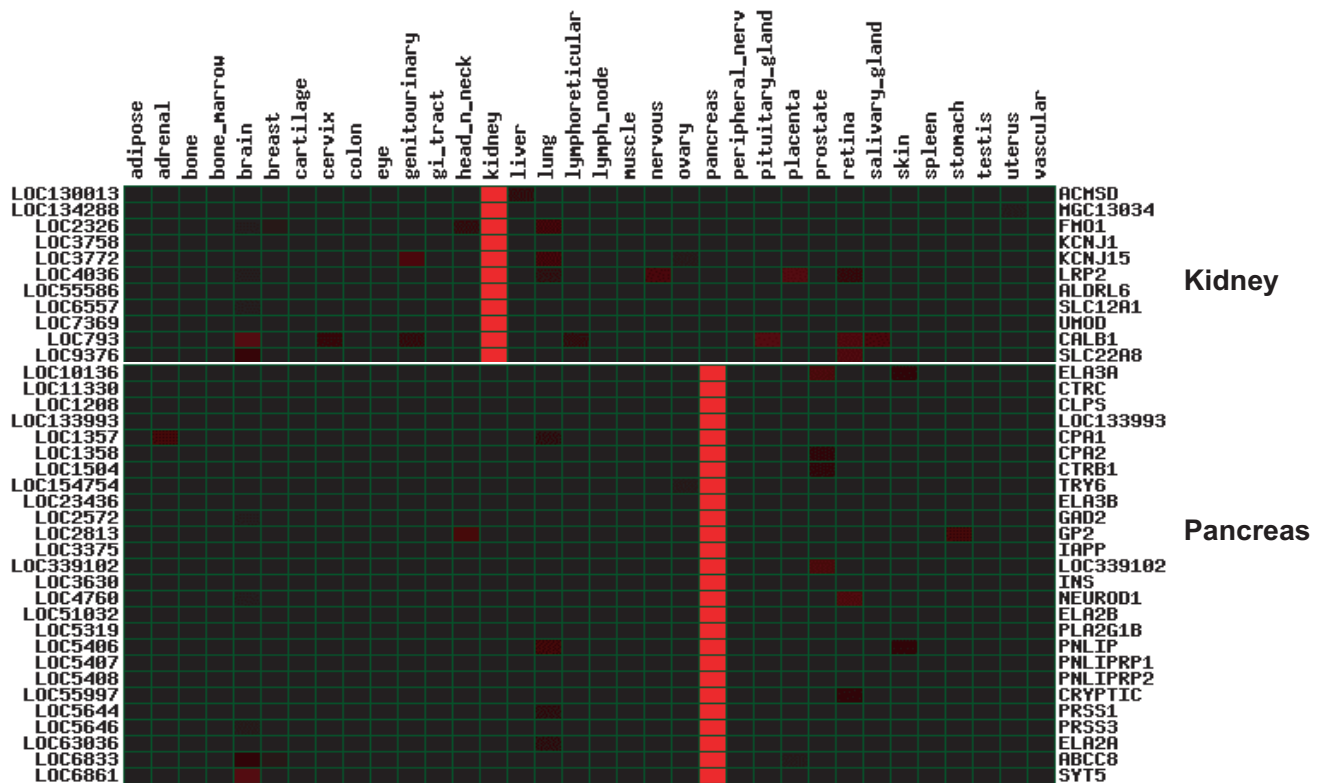


Fig. 3. Heat-map representation of expression patterns of kidney- and pancreas-specific genes found from the GEPIS tissue-specificity screen. Each row represents one distinct gene and each column one type of normal tissue. The LocusLink ID of each gene is shown to the left and gene symbol to the right. Each cell is colored according to its DEU value in normal tissues and is set with an RGB vector $(r, 0, 0)$, where r is the integer value of DEU. Cells with DEU value of zero are therefore shown in black. DEU values between 0 and 255 are mapped to the primary red color with increasing intensity. DEU values which are equal to or greater than 255 are shown as bright red with RGB vector $(255, 0, 0)$. The complete heat-map that includes all the 240 tumor-specific genes is shown in Supplementary Figure 1.

of the *MMP* gene family along the chromosome but also the tissue-specific nature of various *MMP* genes. In addition, it reveals that most of the genes in this region exhibit higher expression in colon tumor compared with normal colon, suggesting a possible link to regional aberrant DNA amplification in colon cancer. Furthermore, the consistent high expression of a neighboring hypothetical gene (*MGC2714*) in tumor tissues implies a possible tumorigenic role for this gene, like the rest of the *MMP* family members.

DISCUSSION

Although the concept of using ESTs for expression analysis has been known for over a decade, the full value of EST-based expression has been somewhat overshadowed by the emerging microarray technology. While microarrays can be an extremely powerful method, EST data can be a strong alternative or even an advantageous method in some cases. Depending on the nature of specific studies, ESTs potentially offer several advantages. First, the extensive gene coverage by ESTs (Brentani *et al.*, 2003) allows expression analysis of almost all

genes, and gene representation is not affected by the sequence picking and designing process required by microarray chips. Furthermore, all the raw EST and tissue source data are already publicly available; therefore, much of the laborious laboratory work can be bypassed when studying the expression in many diverse tissues. Moreover, since the expression level is represented by mRNA abundance over total transcripts and is independent of probe selection and hybridization intensity, EST-based analysis can be a more quantitative and direct measurement of gene expression than microarray-based results. In fact, based on the EST data, we were able to perform large-scale screening for therapeutic tumor antigen targets, namely cell surface proteins with high and specific expression in tumor cells, and many of these targets have been experimentally validated by a variety of methods including tissue microarray, *in situ* hybridization and quantitative reverse transcription-PCR RT-PCR (data not shown).

For ESTs to be widely used and accepted for expression analysis, appropriate tools need to be in place to integrate tissue, EST and genomic information, and to provide reliable expression results that can be easily interpreted. Some tools

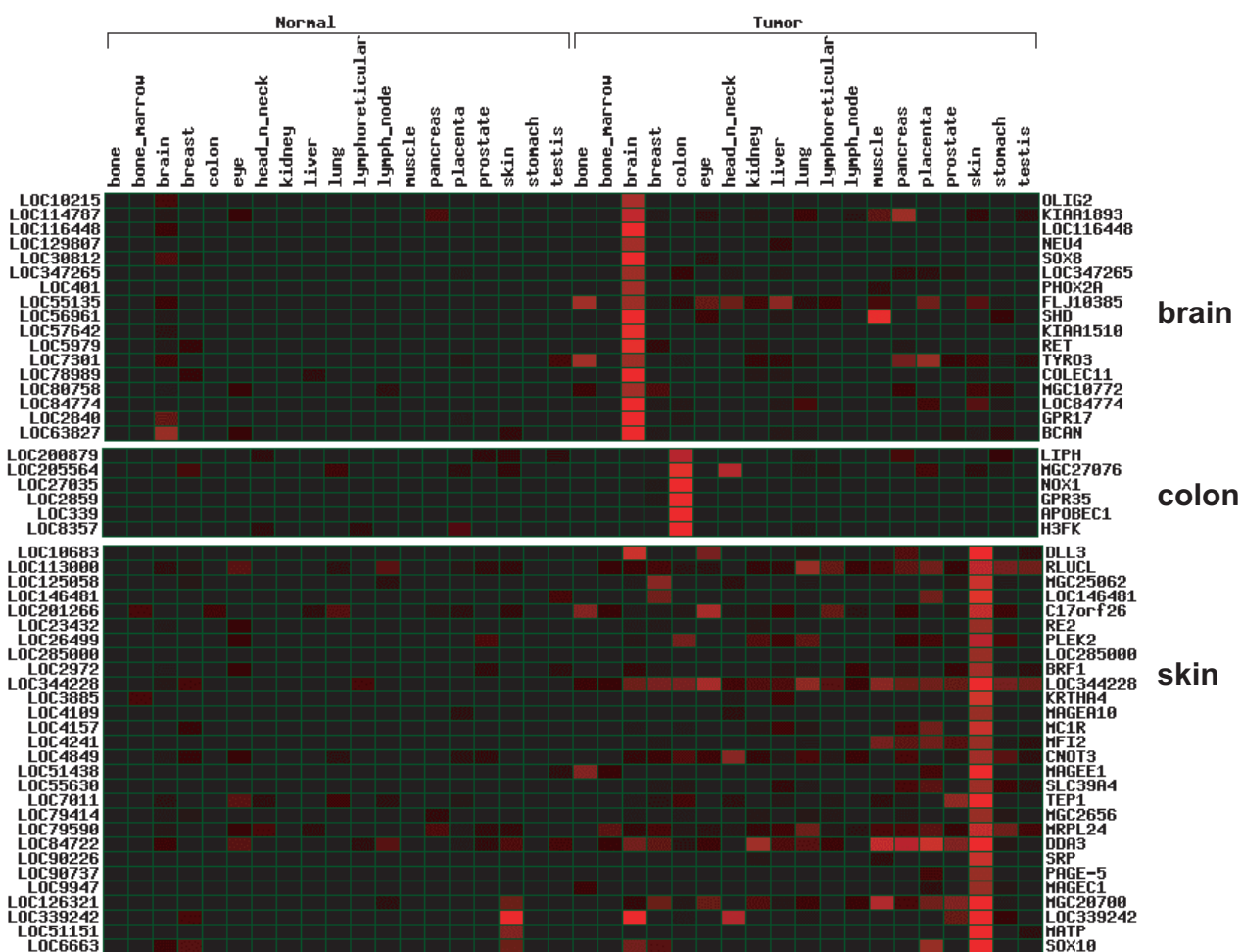


Fig. 4. Heat-map representation of tumor/normal expression patterns of groups of genes found from the GEPIS tumor-specificity screen. This figure is a compilation of three groups of genes that appear to be specific to brain-, colon- and skin-tumor tissues. Each row again represents one distinct gene, and its LocusLink ID is shown to the left and gene symbol to the right. A panel of normal tissues is included on the left and a panel of the corresponding tumor tissues is included on the right. The coloring RGB scheme is identical to that described in Figure 3. The complete heat-map that includes all the 370 tumor-specific genes is shown in Supplementary Figure 2.

have been built for EST-based comparative expression analysis, such as the digital differential display (DDD) (Scheurle *et al.*, 2000) available at NCBI and cDNA xProfiler at CGAP. SAGEmap (Lash *et al.*, 2000), another useful gene expression tool, is based on SAGE data instead of ESTs but is nonetheless similar in principle to EST-based expression analysis. The GEPIS program described here complements existing tools with its ability to display the expression pattern over a large number of normal and tumor samples for individual genes as well as the expression characteristics of a group of genes in a genomic region. In addition to rendering the graphical displays, GEPIS offers a way to rapidly screen for genes with desirable expression patterns, e.g. specific expression in prostate cancer, by modifying a few settings of the program. Therefore, we are providing all the source code files so that they can be adapted for different purposes. For

example, the screens for tissue-specific and tumor-specific genes described here can be easily modified to suit more stringent or relaxed purposes by adjusting the *P*-value requirement and fold differences.

The reliability of EST-based expression results should be ultimately validated by experimental methods. For the genes we have tested so far, we observed a good correlation between EST results and data from tissue microarray and quantitative RT-PCR. The success in making GEPIS reliable can be partially attributed to our data cleansing efforts, where problematic tissue libraries (subtracted or normalized) were excluded, questionable libraries avoided and mislabeled libraries corrected. It is also important to aggregate individual EST libraries at an appropriate level. While more detailed classification of libraries allows expression analysis in more specific tissue types (e.g. infiltrating duct carcinoma versus

breast cancer), sample pooling enriches data content and thereby makes results more reliable. We are confident with the expression results when the combined library size is over 10 000.

It is worth noting that ESTs from other organisms could in principle be used by the GEPIS application for expression analysis. The reliability of EST-based expression results is expectedly dependent on the number of available ESTs. In particular, the four-million mouse ESTs in the NCBI dbEST database represent another attractive data source for expression studies. Preliminary analysis of mouse ESTs indicates a strong bias toward normal mouse tissues. Although the low mouse cancer EST content precludes reliable cancer transcriptome analysis at this stage, the remaining normal ESTs could be excellent for normal tissue expression profiling. As ESTs continue to accumulate in the public domain, the quality of GEPIS analysis should improve further. The availability of this tool will enhance the functional understanding of genes as well as target identification for therapeutic purposes.

ACKNOWLEDGEMENT

We thank Allison Waugh for critical review and comments.

REFERENCES

- Adams,M.D., Kelley,J.M., Gocayne,J.D., Dubnick,M., Polymeropoulos,M.H., Xiao,H., Merril,C.R., Wu,A., Olde,B. and Moreno,R.F. (1991) Complementary DNA sequencing: expressed sequence tags and human genome project. *Science*, **252**, 1651–1656.
- 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. Genet.*, **4**, 256–267.
- Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
- Audic,S. and Claverie,J.M. (1997) The significance of digital gene expression profiles. *Genome Res.*, **7**, 986–995.
- Basarab,T., Picard,J.K., Simpson,E. and Russell-Jones,R. (1999) Melanoma antigen-encoding gene expression in melanocytic naevi and cutaneous malignant melanomas. *Br. J. Dermatol.*, **140**, 106–108.
- Boguski,M.S., Lowe,T.M. and Tolstoshev,C.M. (1993) dbEST—database for ‘expressed sequence tags’. *Nat. Genet.*, **4**, 332–333.
- Brasseur,F., Rimoldi,D., Lienard,D., Lethe,B., Carrel,S., Arienti,F., Suter,L., Vanwijck,R., Bourlond,A., Humblet,Y. *et al.* (1995) Expression of MAGE genes in primary and metastatic cutaneous melanoma. *Int. J. Cancer*, **63**, 375–380.
- 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., Hide,W. *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.
- Chen,P.W., Murray,T.G., Salgaller,M.L. and Ksander,B.R. (1997) Expression of MAGE genes in ocular melanoma cell lines. *J. Immunother.*, **20**, 265–275.
- Clements,J., Hooper,J., Dong,Y. and Harvey,T. (2001) The expanded human kallikrein (KLK) gene family: genomic organisation, tissue-specific expression and potential functions. *Biol. Chem.*, **382**, 5–14.
- Diamandis,E.P., Yousef,G.M., Luo,L.Y., Magklara,A. and Obiezu,C.V. (2000) The new human kallikrein gene family: implications in carcinogenesis. *Trends Endocrinol. Metab.*, **11**, 54–60.
- Ewing,R.M., Kahla,A.B., Poirot,O., Lopez,F., Audic,S. and Claverie,J.M. (1999) Large-scale statistical analyses of rice ESTs reveal correlated patterns of gene expression. *Genome Res.*, **9**, 950–959.
- Gary,S.C., Zerillo,C.A., Chiang,V.L., Gaw,J.U., Gray,G. and Hockfield,S. (2000) cDNA cloning, chromosomal localization, and expression analysis of human BEHAB/brevican, a brain specific proteoglycan regulated during cortical development and in glioma. *Gene*, **256**, 139–147.
- Green,M.C., Murray,J.L. and Hortobagyi,G.N. (2000) Monoclonal antibody therapy for solid tumors. *Cancer Treat. Rev.*, **26**, 269–286.
- Heid,C.A., Stevens,J., Livak,K.J. and Williams,P.M. (1996) Real time quantitative PCR. *Genome Res.*, **6**, 986–994.
- Hishiki,T., Kawamoto,S., Morishita,S. and Okubo,K. (2000) BodyMap: a human and mouse gene expression database. *Nucleic Acids Res.*, **28**, 136–138.
- Jaworski,D.M., Kelly,G.M., Piepmeier,J.M. and Hockfield,S. (1996) BEHAB (brain enriched hyaluronan binding) is expressed in surgical samples of glioma and in intracranial grafts of invasive glioma cell lines. *Cancer Res.*, **56**, 2293–2298.
- Lash,A.E., Tolstoshev,C.M., Wagner,L., Schuler,G.D., Strausberg,R.L., Riggins,G.J. and Altschul,S.F. (2000) SAGEmap: a public gene expression resource. *Genome Res.*, **10**, 1051–1060.
- Louis,T. (1981) Confidence intervals for a binomial parameter after observing no success. *Am. Stat.*, **35**, 154.
- Lucas,S., De Smet,C., Arden,K.C., Viars,C.S., Lethe,B., Lurquin,C. and Boon,T. (1998) Identification of a new MAGE gene with tumor-specific expression by representational difference analysis. *Cancer Res.*, **58**, 743–752.
- Pruitt,K.D. and Maglott,D.R. (2001) RefSeq and LocusLink: NCBI gene-centered resources. *Nucleic Acids Res.*, **29**, 137–140.
- Ramaswamy,S., Tamayo,P., Rifkin,R., Mukherjee,S., Yeang,C.H., Angelo,M., Ladd,C., Reich,M., Latulippe,E., Mesirov,J.P. *et al.* (2001) Multiclass cancer diagnosis using tumor gene expression signatures. *Proc. Natl Acad. Sci., USA*, **98**, 15149–15154.
- Ross,D.T., Scherf,U., Eisen,M.B., Perou,C.M., Rees,C., Spellman,P., Iyer,V., Jeffrey,S.S., Van de Rijn,M., Waltham,M. *et al.* (2000) Systematic variation in gene expression patterns in human cancer cell lines. *Nat. Genet.*, **24**, 227–235.
- Scheurle,D., DeYoung,M.P., Binniger,D.M., Page,H., Jahanzeb,M. and Narayanan,R. (2000) Cancer gene discovery using digital differential display. *Cancer Res.*, **60**, 4037–4043.
- Schmitt,A.O., Specht,T., Beckmann,G., Dahl,E., Pilarsky,C.P., Hinzmann,B. and Rosenthal,A. (1999) Exhaustive mining of EST libraries for genes differentially expressed in normal and tumour tissues. *Nucleic Acids Res.*, **27**, 4251–4260.
- Spellman,P.T. and Rubin,G.M. (2002) Evidence for large domains of similarly expressed genes in the Drosophila genome. *J. Biol.*, **1**, 5.

- Su,A.I., Cooke,M.P., Ching,K.A., Hakak,Y., Walker,J.R., Wiltshire,T., Orth,A.P., Vega,R.G., Sapinoso,L.M., Moqrich,A. et al. (2002) Large-scale analysis of the human and mouse transcriptomes. *Proc. Natl Acad. Sci., USA*, **99**, 4465–4470.
- Viapiano,M.S., Matthews,R.T. and Hockfield,S. (2003) A novel membrane-associated glycovariant of BEHAB/brevican is up-regulated during rat brain development and in a rat model of invasive glioma. *J. Biol. Chem.*, **278**, 33239–33247.
- Zhou,Y., Luoh,S.M., Zhang,Y., Watanabe,C., Wu,T.D., Ostland,M., Wood,W.I. and Zhang,Z. (2003) Genome-wide identification of chromosomal regions of increased tumor expression by transcriptome analysis. *Cancer Res.*, **63**, 5781–5784.