Gene expression

Quantitating tissue specificity of human genes to facilitate biomarker discovery

George Vasmatzis1,*,†, Eric W. Klee1,*,†, Dagmar M. Kube1, Terry M. Therneau2 and Farhad Kosari1

1Mayo Clinic Comprehensive Cancer Center and Division of Experimental Pathology, Department of Laboratory Medicine and Pathology, 200 First St. SW, Rochester, MN 55905, USA and 2Division of Biostatistics, Health Sciences Research, Mayo Clinic, USA

Received on July 11, 2006; revised on January 16, 2007; accepted on March 10, 2007
Advance Access publication March 23, 2007
Associate Editor: Alfonso Valencia

ABSTRACT
We describe a method to identify candidate cancer biomarkers by analyzing numeric approximations of tissue specificity of human genes. These approximations were calculated by analyzing predicted tissue expression distributions of genes derived from mapping expressed sequence tags (ESTs) to the human genome sequence using a binary indexing algorithm. Tissue-specificity values facilitated high-throughput analysis of the human genes and enabled the identification of genes highly specific to different tissues. Tissue expression distributions for several genes were compared to estimates obtained from other public gene expression datasets and experimentally validated using quantitative RT-PCR on RNA isolated from several human tissues. Our results demonstrate that most human genes (~98%) are expressed in many tissues (low specificity), and only a small number of genes possess very specific tissue expression profiles. These genes comprise a rich dataset from which novel therapeutic targets and novel diagnostic serum biomarkers may be selected.

Contact: vasm@mayo.edu
Supplementary information: Supplementary data are available at Bioinformatics online.

1 INTRODUCTION

The early detection of cancer through general screening assays is known to greatly impact patient care and outcome. However, most known biomarkers lack sufficient specificity and sensitivity of disease detection to provide adequate predictive value in general population tests (Benowitz, 2004). Diminished assay performance is often attributable to the complexity of measuring analyte levels in serum due to low tissue specificity of the targeted gene product. The tissue specificity of a gene is a measure of the relative distribution of gene expression across major tissue types in the human body. Tissue-specific genes enable an assay to detect small increases in

*To whom correspondence should be addressed.
†The authors wish it to be known that, in their opinion, the first two authors should regarded as joint First Authors.
gene expression values derived from EST data. Relative level of gene expression in a tissue can be approximated by counting ESTs corresponding to that gene (Vasmatzis et al., 1998). The EST sequences are mapped to the human genome with an adapted binary indexing algorithm (see Supplementary Material) which stores a binary representation of the complete human genome sequence (Kent, 2002) in RAM. Binary representations of EST sequences are then rapidly mapped to the binary encoded genome through direct memory access. This approach drastically increases the efficiency of the mapping process by avoiding the computational overhead of traditional search algorithms. Additionally, this method is wholly independent of the UniGene EST cluster assembly protocol and avoids many of the shortcomings associated with traditional clustering algorithms, such as the formation of superclusters by groups of highly homologous sequences (Vasmatzis et al., 1998). We have previously described a similar method for assessing tissue specificity and used it to mine prostate EST sequence sets and identify Cysteine-rich secretory protein-3 (CRISP-3) as a highly specific biomarker in prostate adenocarcinoma (Asmann et al., 2002).

The tissue specificity of several genes in many tissue types is reported and these scores used to comprehensively analyze the EST dataset. Tissue specificity is calculated by analyzing the EST sequences mapped to a target genome segment and then differentiating the ESTs by cDNA library tissue type and disease state (cancer or normal). Using this numeric value, a subset of highly tissue specific genes was identified. The tissue-specificity profiles for several genes were independently evaluated using quantitative PCR and compared to tissue-specificity data obtained from other methods. This study describes a novel method for quantitating tissue specificity in normal and cancerous tissues and augmenting high-throughput biomarker discovery projects.

2 METHODS

2.1 EST sequences

7 597 006 Human EST sequences were downloaded from NCBI’s dbEST on September 03, 2006 and stored as FASTA sequences with the identification tag containing the EST identifier, library ID, dbEST ID, putative gene ID, clone ID, nucleotide index where quality sequence ends, and when available the partner EST name. EST sequences were classified by annotations on tissue/organ type (brain, prostate, ovary, etc.), tissue histology (cancer, normal) and cell extraction method (bulk, micro-dissected, cell line) abstracted from 8 605 originating cDNA libraries. cDNA library descriptions were obtained using the Gene Library Summarizer at CGAP (http://cgap.nci.nih.gov/Tissues/LibrarySummarizer and downloading the Hs_LibData.dat file).

2.2 Genomic and mRNA sequences

The human genomic sequence was downloaded in March 2006 from NCBI (ftp.ncbi.nih.gov/genomes/H_sapiens/CHR_+) and served as the template against which EST and mRNA sequences are mapped. RefSeq Human mRNA sequences were downloaded from NCBI (ftp.ncbi.nih.gov/genomes/H_sapiens/RNA/rna.fa.gz) and augmented with the German Human Genome Project (NGFN) complete human cDNA sequence set, downloaded from GenBank. These sequences were used to better characterize the position of genes on the human genome.

2.3 Binary indexing algorithm (see Supplementary Material for more details)

This algorithm converts the human genome sequence into consecutive 32-letter words, where each word consists of two 32 bit binary numbers. The first number (the base) encodes A’s and G’s as 1’s and T’s and C’s as 0’s, and the second number (the check) encodes A’s and T’s as 0’s and G’s and C’s as 1’s. For each word, the base number, the check number, the position of the word in the genomic sequence, and the sequence identifier are written to a file. A storage routine then stores the binary form of the human genome in RAM, using external chains to track the position of the sequentially ordered base numbers.

2.4 Mapping ESTs and mRNAs (see Supplementary Material for more details)

An EST or mRNA nucleotide sequence is mapped to the binary human genome by creating a series of overlapping 32-letter words spanning each sequence and its reverse compliment. Base and check binary numbers are created for all words using the same methods used to encode the human genomic sequences. An EST base number acts as an index to directly identify the beginning of the corresponding genomic sequence external chain. The EST query number is then cross-referenced against the check number. If check numbers do not match, the next position of the EST base number matches evaluated. This process is repeated until the whole chain is exhausted, for all 32-bit binary numbers (forward and reverse compliment sequences). When the algorithm finds a hit, the identifier and position are stored in temporary arrays. When all sequence information is processed, the information in the temporary arrays is used to combine hits and identify large regions of overlap between the EST query sequence and the human genome. Due to extensive homology in the genome, many sequences possess hits with a high degree of overlap. Therefore, only the segment in the indexed database with the largest number of consecutive hits is selected. In cases where a query sequence maps to more than one location on the genome, the segment of the genome with more exon/intron boundaries is given priority. This is done to distinguish a gene from a pseudogene(s), which often does not have introns. Each segment is then processed by a routine to more rigorously define the first and last nucleotide positions of the overlapping region between the query sequence and the genomic sequence.

2.5 Numeric index of tissue specificity

Using the previously described algorithm, results from mapping EST sequences are summarized in two matrices, one for cancer and one for normal. Each matrix has one row per EST cluster and one column per tissue. There were 47 440 unique EST clusters in 53 different tissues. For an array define:

\[ y_{ij} = \text{EST count for EST cluster } i \text{ and tissue } j \]

\[ g_i = \sum y_{ij} = \text{total EST count for EST cluster } i \]

\[ t_j = \sum y_{ij} = \text{total EST count for tissue } j \]

\[ N = \sum t_j = \text{total EST count for the matrix} \]

The natural estimate of expression for EST cluster \( i \) in tissue \( j \) is the proportion \( p_{ij} = y_{ij} / t_j \). As an index of specificity, we propose the ratio between the EST cluster expression within a particular tissue to the total sum for the EST cluster.

\[ S_{ij} = \frac{p_{ij}}{\sum_k p_{ik}} \]
This index can be computed for only cancer tissues, only normal tissues, or for a combination of samples by creating the appropriate matrix for $\gamma$. An overall $P$-value for each EST cluster, to control for the problem of multiple comparisons of tissues within an EST cluster, is computed using the omnibus test for the fit of a Poisson model. This model is essentially equivalent to a binomial model due to the low values of $\rho$, all of which are <0.03 (McCullagh and Nelder, 1983).

$$C_i = \sum_j y_{ij} \log \left( \frac{y_{ij}}{\hat{p}_i} \right)$$

Where, $\hat{p}_i$ is the overall prevalence of EST cluster $i$ over all EST sequences in the data matrix. Under the null hypothesis, non-differential expression for EST cluster $i$ across tissues, $C_i$ will follow a chi-squared distribution on 53 degrees of freedom. A cutoff of $p < 0.01$ corresponds to $C > 79.8$, and $p < 0.001$ corresponds to $C > 90.6$, for instance.

The above maximum likelihood estimates can be unstable for low expressing EST. We used a Bayes shrinkage estimate, which shrinks outliers back towards a common mean:

$$p^*_i = \frac{y_i + a_i}{t_i + \hat{p}}$$

This is based on a beta-binomial prior, a fairly standard approach, with

$$a_i = \frac{(1 - \hat{p})}{CV^2} - \hat{p}_i$$

$$\hat{p}_i = \frac{a_i (1 - \hat{p}_i)}{\hat{p}}$$

where $\hat{p}$ is the overall average expression for the EST, and CV is the coefficient of variation of the true $p_i$ values about this center. The optimal Bayes estimate occurs for the true (but unknown) CV value; $CV = 0$ will force $p^*_i = \hat{p}$ for all tissues, i.e. no differential expression, and $CV = \infty$ gives the unconstrained values $p_i$. We chose to use a constant value of $CV = 3$, which corresponds to a modest shrinkage. Any chosen value gives a tradeoff between sensitivity and specificity: larger choices for CV produce, more ‘differentially expressed, ESTs, particularly those of tissues, or for a combination of samples by creating the appropriate matrix for $\gamma$. An overall $P$-value for each EST cluster, to control for the problem of multiple comparisons of tissues within an EST cluster, is computed using the omnibus test for the fit of a Poisson model. This model is essentially equivalent to a binomial model due to the low values of $\rho$, all of which are <0.03 (McCullagh and Nelder, 1983).

$$C_i = \sum_j y_{ij} \log \left( \frac{y_{ij}}{\hat{p}_i} \right)$$

Where, $\hat{p}_i$ is the overall prevalence of EST cluster $i$ over all EST sequences in the data matrix. Under the null hypothesis, non-differential expression for EST cluster $i$ across tissues, $C_i$ will follow a chi-squared distribution on 53 degrees of freedom. A cutoff of $p < 0.01$ corresponds to $C > 79.8$, and $p < 0.001$ corresponds to $C > 90.6$, for instance.

The above maximum likelihood estimates can be unstable for low expressing EST. We used a Bayes shrinkage estimate, which shrinks outliers back towards a common mean:

$$p^*_i = \frac{y_i + a_i}{t_i + \hat{p}}$$

This is based on a beta-binomial prior, a fairly standard approach, with

$$a_i = \frac{(1 - \hat{p})}{CV^2} - \hat{p}_i$$

$$\hat{p}_i = \frac{a_i (1 - \hat{p}_i)}{\hat{p}}$$

where $\hat{p}$ is the overall average expression for the EST, and CV is the coefficient of variation of the true $p_i$ values about this center. The optimal Bayes estimate occurs for the true (but unknown) CV value; $CV = 0$ will force $p^*_i = \hat{p}$ for all tissues, i.e. no differential expression, and $CV = \infty$ gives the unconstrained values $p_i$. We chose to use a constant value of $CV = 3$, which corresponds to a modest shrinkage. Any chosen value gives a tradeoff between sensitivity and specificity: larger choices for CV produce, more ‘differentially expressed, ESTs, particularly those of low overall abundance, smaller values for CV give a shorter list.

2.6 Comparative tissue specificity analysis

Tissue-specificity profiles were obtained via the webinterfaces of UniGene’s EST ProfileViewer, CGAP’s SAGE Genie, the Ludwig Institute’s MPSS Database and the Novartis Research Institute’s SymAtlas v.1.1.1. All queries were performed using default settings for human sequences. Comparisons were made for the normal tissues common to all five techniques. These tissues include: heart, kidney, placenta, bone marrow, colon, prostate, pancreas, uterus and lung. Agreement between methods on identifying the tissue/organ having the highest expression of a target gene was measured. Additionally, observations were made on how well the methods agreed on the relative expression values and normalized them so that expression value for each gene across all tissue types varied from zero to five.

### Table 1. List of 22 human genes that were analyzed by Q-RT-PCR

<table>
<thead>
<tr>
<th>GENE ID</th>
<th>REFSEQ ID</th>
<th>Gene name</th>
</tr>
</thead>
<tbody>
<tr>
<td>ENPEP</td>
<td>NM_001977</td>
<td>Glutamyl aminopeptidase</td>
</tr>
<tr>
<td>SLC2A9</td>
<td>NM_020041</td>
<td>Solute carrier family</td>
</tr>
<tr>
<td></td>
<td></td>
<td>facilitated glucose</td>
</tr>
<tr>
<td></td>
<td></td>
<td>transporter member 9</td>
</tr>
<tr>
<td>SLC12A1</td>
<td>NM_000338</td>
<td>Solute carrier</td>
</tr>
<tr>
<td>UMOD</td>
<td>NM_003361</td>
<td>Uromodulin precursor</td>
</tr>
<tr>
<td>CUBN</td>
<td>NM_001081</td>
<td>Cubulin</td>
</tr>
<tr>
<td>SLC13A3</td>
<td>NM_022829</td>
<td>Solute carrier family</td>
</tr>
<tr>
<td>TMEM27</td>
<td>NM_020665</td>
<td>Homo sapiens</td>
</tr>
<tr>
<td></td>
<td></td>
<td>transmembrane protein</td>
</tr>
<tr>
<td>CLDN2</td>
<td>NM_020384</td>
<td>Claudin-2</td>
</tr>
<tr>
<td>KLOTHO</td>
<td>NM_004795</td>
<td>Klotho, transcript variant 1</td>
</tr>
<tr>
<td>TMRSS2</td>
<td>NM_005656</td>
<td>Transmembrane</td>
</tr>
<tr>
<td></td>
<td></td>
<td>protease, serine 2</td>
</tr>
<tr>
<td>Protein</td>
<td>NM_031102</td>
<td>Prostate protein</td>
</tr>
<tr>
<td>PSA</td>
<td>NM_001648</td>
<td>Prostate-specific antigen</td>
</tr>
<tr>
<td>ACPP</td>
<td>NM_001099</td>
<td>Prostatic acid phosphatase</td>
</tr>
<tr>
<td>SEMG1</td>
<td>NM_003007</td>
<td>Semenogelin 1</td>
</tr>
<tr>
<td>KLK2</td>
<td>NM_005551</td>
<td>Kallikrein 2 precursor</td>
</tr>
<tr>
<td>MSMB</td>
<td>NM_002443</td>
<td>Homo sapiens</td>
</tr>
<tr>
<td></td>
<td></td>
<td>beta-microsominoprotein</td>
</tr>
<tr>
<td></td>
<td></td>
<td>transcript variant PSP94</td>
</tr>
<tr>
<td>CRISP3</td>
<td>NM_006061</td>
<td>Cysteine-rich secretory protein 3</td>
</tr>
<tr>
<td>SFTPA1</td>
<td>NM_005411</td>
<td>Surfactant pulmonary-associated</td>
</tr>
<tr>
<td>SFTPB</td>
<td>NM_000542</td>
<td>protein A</td>
</tr>
<tr>
<td>UGR1</td>
<td>NM_054023</td>
<td>Uteroglobin-related protein 1</td>
</tr>
<tr>
<td>NAPSA</td>
<td>NM_004851</td>
<td>Homo sapiens</td>
</tr>
<tr>
<td>GAPDH</td>
<td>M33197</td>
<td>Glyceraldehyde-3-phosphate</td>
</tr>
<tr>
<td></td>
<td></td>
<td>dehydrogenase</td>
</tr>
</tbody>
</table>

2.7 Clustering

The maximum tissue-specificity values for the top 140 genes were plotted by using Visual Basic in Microsoft Excel. The hierarchical map of qPCR expression values of genes was constructed using Cluster version 3.0 (Eberspecher et al., 1995) and visualized using Java TreeView version 1.0.8 (http://jtreeview.sourceforge.net).

2.8 Quantitative PCR

Panels of normalized, first-strand cDNA preparations, Human MTC Panel I and II cDNAs (BD Biosciences Clontech, Palo Alto, CA, USA) were used to determine the relative expression levels of specific transcripts in 16 different human tissues/cells. Equal amounts of each cDNA were analyzed by quantitative PCR using gene-specific primers and SYBR Green PCR Master Mix (Applied Biosystems, Foster City, CA, USA) on an ABI PRISM 7900HT Sequence Detection System. Expression levels for the 22 human genes listed in Table 1 were analyzed simultaneously in a 384-well plate. We converted the cycle threshold (Ct) for each gene to relative expression values and normalized them so that expression value for each gene across all tissue types varied from zero to five.

Primers specific for the housekeeping gene GAPDH were used to verify cDNA quality/reaction set-up effectiveness as recommended in the cDNA Panels User Manual (Protocol # PT3158-1, Version # PR35997, 21 May 2003, BD Biosciences Clontech).

3 RESULTS

3.1 Identification of the most specific human EST clusters

Tissue-specificity values were computed for 9385 human EST clusters with tissue-type differential expression at $p < 0.001$, in each tissue-type library, except for the pooled tissue libraries.
labeled as uncharacterized tissue, placenta, pooled tissue and whole body. Values were generated using an EST count matrix consisting of cancer library plus normal library ESTs in the target tissue and normal library ESTs for all other tissues. The maximum tissue-specificity value for an EST cluster was used to rank the complete EST cluster list. The maximum specificity value for each EST cluster, with $P < 0.001$ for the global significance test, is plotted against the total number of EST sequences in each cluster (Fig. 1). Genes specific to a tissue are found on the right side of Figure 1. We suspect high tissue specificity makes these genes a rich set of candidates for serum biomarkers and/or therapeutic targets. To explore this assertion, known and potential biomarkers were plotted and highlighted in Figure 1. Many of these markers have high maximum tissue-specificity values and are over represented, relative to all plotted genes, in the right two quartiles. For example, CEA and PSA had significant differential expression amongst tissue types and have maximum specificity values of 0.52 and 0.69, respectively. These two biomarkers are found in the upper right side of Figure 1, indicating that they are highly tissue specific and highly expressed. Figure 1 clearly shows there are many highly specific genes, found in the highest quartile, that have not been reported as biomarkers that may have valuable diagnostic, prognostic and/or therapeutic purposes. It is interesting to note that CA125, a widely used serum biomarker for ovarian cancer, was not found to be differentially expressed amongst tissue types (at $P < 0.001$) and was therefore non-specifically expressed, demonstrating that not all serum biomarkers necessarily display high-tissue specificity.

A preponderance of the EST clusters have a maximum tissue-specificity value of less than 0.5 and are not specific to any one tissue (Fig. 2). To further investigating gene clusters with high tissue specificity, we arbitrarily chose to examine EST clusters with a maximum tissue specificity $>0.7$. One hundred and forty EST clusters, <1% of the total, were identified, clustered by tissue, and plotted on a heat map (Fig. 3). Of the 49 tissue types identified, 37 have at least one EST cluster for which that tissue has the highest expression; with liver containing more specifically expressed EST clusters than any other tissue/organ in this set.

### 3.2 Evaluation of statistical relevance and false discovery rate

The primary purpose of the $C$ statistic threshold is to avoid spurious reports of high specificity for a gene, particularly for low expressed genes with little representation in the
EST libraries. The P-value from the test correctly accounts for multiple testing between tissues for a single gene, but is not calibrated across all genes. To properly evaluate the performance of the combined selection criteria of P < 0.001 and specificity index > 0.5, we computed the permutation distribution (see Supplementary Material). Specifically, a random dataset was generated 100 times, subject to the row and column constraints of the original data matrix: the total number of entries for any given tissue (columns) and for any given gene (rows) are as observed in the original data. This produces a dataset with no association between gene and tissue, but retaining the global characteristics of both rare and common genes, and heavily and lightly sampled tissues. For each random dataset, the selection procedure was applied, that is, each row of the result was labeled by its most prevalent tissue type, and rows that satisfy the selection criteria were retained. Table 2 shows results for 5 of the 53 tissue types, the full matrix of results may be found in the Supplementary Materials.

In 100 random samples, the maximum tissue specificity observed for a sample which passed the statistical test, was 0.412; thus the estimated FDR for a threshold of 0.5 or 0.7 is 0, for all tissues. Looking at the first tissue in Table 2, adipose, the average number of ESTs from the randomized dataset which passed the statistical threshold was 127, the number which passed the specificity threshold of 0.2 was 43 (not shown), and the number which passed both was 0.04; giving estimated FDR rates of 41 (=127/(127+184)*100%), 28 and 0.1%, respectively. Thus, we see that both selection criteria play an important role. To ensure an FDR based on 100 random samples was sufficient, a second 100 random samples were generated and FDR rates computed based on all 200 random samples. There was no statistical difference in the FDR rates computed from the 200 random samples compared to the FDR rates computed from the initial 100 random samples (see Supplementary Table S-1).

### 3.3 Identification, comparison and validation of candidate cancer biomarkers

Our research interests include serum biomarker discovery in prostate, kidney and lung cancers. Therefore, we chose to identify genes that are highly specific to one of these tissues. The computed specificity indices for these genes were experimentally validated using quantitative PCR on commercially available panels of normalized, first-strand cDNA preparations from 16 different normal human tissues/cells. Hierarchical clustering of the normalized expression values of these genes grouped them according to their gene expression profiles in the 16 normal human tissues/cells. A heat map of the clustered quantitative PCR data is shown in Figure 4. Genes specific to a tissue/cell type clustered together (Fig. 4), as was observed with our computed specificity indices (Fig. 3).

Fig. 2. Histogram of the distribution of human EST clusters according to organ specificity.

Table 2. Sample dataset of the number of gene clusters from the actual EST data and the ones resulted by the permutation test (see Supplementary Material for the complete table)

<table>
<thead>
<tr>
<th>Tissue</th>
<th>Actual data</th>
<th>Averaged randomized data (100 iterations)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Assigned</td>
<td>Cstat &gt;90.6</td>
</tr>
<tr>
<td>Adipose</td>
<td>562</td>
<td>184</td>
</tr>
<tr>
<td>Adrenal cortex</td>
<td>1554</td>
<td>478</td>
</tr>
<tr>
<td>Adrenal medulla</td>
<td>471</td>
<td>187</td>
</tr>
<tr>
<td>Bone</td>
<td>452</td>
<td>159</td>
</tr>
<tr>
<td>Bone marrow</td>
<td>553</td>
<td>207</td>
</tr>
</tbody>
</table>
expressed in lung include: SFTPA1, SFTPB, UGR1 and NAPSA. The difference between the expression level in the lung and the next highest expression level in any other organ ranged from 800- to 2000-fold, for all genes except NAPSA, for which the difference was \(~4\)-fold. The observed variation in GAPDH expression across all of the tissues/cells except skeletal muscle was \(~5\)-fold. The observed variation in GAPDH expression across all of the tissues/cells including skeletal muscle was \(~20\)-fold. The level of GAPDH in skeletal muscle was expected to be exceptionally high, possibly due to the high ATP production in this tissue (cDNA Panels User Manual). The BD Biosciences Clontech normalization procedure for the MTC Panel cDNAs uses the expression levels of several housekeeping genes (\(\alpha\)-tubulin, \(\beta\)-actin, glyceraldehyde-3-phosphate dehydrogenase, and phospholipase A2) to standardize the amount of cDNA present within each MTC Panel (cDNA Panels User Manual). However, the observed 5-fold variation in GAPDH expression across tissues, excluding the aberrantly high expression in skeletal muscle, is significantly higher than the 20% or lower variation in housekeeping gene expression claimed in the cDNA Panels User Manual. This is likely due to significantly higher sensitivity of quantitative PCR relative to traditional semi-quantitative PCR methods. Thus, the observed variation in GAPDH expression across tissues/cells indicates either that the MTC Panel cDNAs were not normalized accurately or that GAPDH is not the best normalization gene.

Tissue expression profiles generated by MPSS, SAGE, UniGene Electronic Northern, SymAtlas and our method showed moderate agreement. All five analysis methods identified the same maximum expression tissue/organ in 8 of 12 genes tested. Two of 12 genes showed agreement between four methods, and 2 of 12 genes showed agreement between three methods. The agreement in percent of total expression value for the consensus tissue type was variable in the nine tissues examined. Of the eight genes with consensus maximum expression tissue/organ type identification, four of eight showed close agreement with relative expression values from 95.7 to 100% of total expression and standard deviations ranging from 0.2 to 1.8%. The four other consensus genes demonstrated greater variation in the relative expression value of the maximum expression tissue, with values of 38.1–100% and SDs of 12.6–24.4%.
4 DISCUSSION

We have developed a method to estimate the expression distribution of a gene in normal and cancer tissues based on a computational analysis of EST data. To facilitate rapid and computationally inexpensive mapping of EST sequences to the human genome, we used an adapted binary indexing approach. This novel method converted the entire human genome sequence information to binary values and stored it in RAM. Binary converted EST sequences were then efficiently mapped to the genome to create gene profiles. We hypothesized that the more a gene is expressed in a tissue, the more it would be represented by clones in the cDNA library and, hence, by ESTs in the corresponding EST library. Therefore, we assumed that the count of ESTs corresponding to a gene from a cDNA library correlates with the expression level of that gene in the tissue from which the library was derived. However, the total number of ESTs in individual libraries is often small, resulting in inaccurate estimates of expression in low abundance genes.

To improve specificity estimates, we combined EST libraries derived from the same tissue, only differentiating libraries into normal and cancer tissue sets (cancer and pre-cancer were pooled together). We assumed that the expression level of a gene in a tissue correlates with the count of ESTs corresponding to the gene in cDNA libraries derived from the tissue. To identify potential biomarkers, we sum the normal and cancer EST counts for each gene in the tissue of interest. The purpose here was to increase the likelihood that genes specific to either cancer or normal tissues are selected.

EST counts cannot be used directly to compare expression of a gene in different tissues because cDNA libraries generated from different tissues have been sampled to different extents. For example, more ESTs have been sequenced from brain tissue than from adipose tissue. Therefore, counts of ESTs corresponding to particular genes were normalized by the total number of ESTs in each tissue. Relative expression levels of a gene in different tissues were then obtained by comparing normalized expression levels of the gene in all tissues. The likelihood of identifying a gene with tissue-specific expression properties is measured by the omnibus test for the fit of a Poisson model and allows a significance level to be associated with the tissue-specificity values calculated.

Known cancer biomarkers were found among the most specific genes (Fig. 1), demonstrating the value and validity of the algorithm presented here. Figures 1 and 2 clearly illustrate that most genes lack any significant tissue specificity. Consequently, a tissue-specificity filter can be expected to significantly reduce the number of candidate genes that are to be evaluated as putative cancer biomarkers. Figure 3 presents a list of highly specific genes that may serve as a valuable resource for the discovery of novel cancer biomarkers and targets.

While most known biomarkers appeared tissue specific, CA125 was found to be insignificant in our analysis and no specificity value computed. In this case, lack of tissue differential expression is not necessarily inconsistent with the gene’s utility as a biomarker. Although, CA125 is expressed in normal and tumor cells (Hardardottir et al., 1990; Nap et al., 1996; O’Brien et al., 1986; Zurawski et al., 1988), cell-surface expression and secretion of CA125 into the extracellular space (Lloyd and Yin, 2001) appear to be associated with the conversion from benign to cancer cells (Meyer and Rustin, 2000). Consistent with this, CA125 has been shown to accumulate in the serum of cancer patients bearing ovarian and other carcinomas (Bast et al., 1983; Bon et al., 1996). Most other secretory protein biomarkers in Figure 1 are not modified in this way and appear to be tissue specific. However, CA125 demonstrates that while tissue specificity can be used as a filter to enrich gene pools for putative biomarkers, non-specific proteins cannot be wholly dismissed. Proteins possessing desirable biomarker characteristics but lacking significant tissue specificity should still be scrutinized for characteristics of the serum form that are unique to the disease state of interest.

Using 12 genes selected for experimental validation, we compared the results from our method to the tissue distribution profiles generated by MPSS, SAGE, UniGene Electronic Northern and SymAtlas. In 83.3% of the genes evaluated (8 of 12), our method identified the same maximal expressed tissue as the other methods, and in 80% of these cases the majority was actually a consensus. For two genes, the maximal expressed tissue identified by our method deviated from the majority prediction. Unfortunately, we were unable to experimentally confirm or refute the results of our method, as the maximal tissue identified by our method was one of two tissues in which qrtPCR validation was not performed (the tissue panel used did not have uterus or bone marrow samples). This limited comparison is not sufficient to precisely compare and contrast the different methods for quantifying tissue expression profiles. But, the agreement between methods, in identifying the maximal expressed tissue for most genes, suggests all five methods are proximally correct. Whereas the variability between the five methods in quantifying the relative expression values for these genes across the nine selected tissues suggests there may be validity in utilizing multiple methods when computing tissue specificity. Coalescing output from independent methods may provide a more complete measure of tissue specificity than any one approach. The exact measure of value added by any single method of ascertaining tissue specificity cannot be determined without a robust comparison of platforms. However, it seems likely that the most accurate picture of tissue specificity will arise from the combination of multiple algorithms into a meta-server type system. Zhang et al. (2004) reported a new system, GEPIIS, which determines organ specificity of genes by performing BLAST searches. The algorithm compares entries in dbEST against the available human gene sequences in LocusLink database and estimates the expression level of genes based on the counts of ESTs corresponding to that gene. This method uses individual t-tests to compare each pair of tissues, separately for each EST cluster. If the counts are greater than 5–10, the Z-statistic is a close approximation to our likelihood statistic C (Anscombe, 1949), but without accounting for multiple testing either between tissues or between EST clusters. Therefore, while there is some overlap between the algorithm presented here and GEPIIS, we feel the approach outlined in this manuscript provide a more appropriate method for computing statistically relevant tissue-specificity values.
Our algorithm assigns the highest specificity index to genes that are highly expressed and tissue specific. Since the algorithm corrects for low expression level and filters for non-specific expression, there may be some biomarkers expressed at a modest level that are more specific than they appear by our algorithm. It is also noteworthy that genes with different tissue expression profiles can have similar specificity indices. For example, a gene whose expression is confined to two different tissues at approximately equal levels could have a specificity index similar to that of another gene expressed predominantly in one tissue, but with lower level of expression across many other tissues. Which tissue expression profile is more desirable depends on the clinical objective and the specifics of the particular profiles. Expression of a gene confined to a female- and male-specific organ, e.g. would not be problematic for a biomarker since these two organs are mutually exclusive in a human body. Of course, the utility of a biomarker also depends on a measurable difference in expression between the normal and disease states. Information about specificity is critical but not sufficient for biomarker discovery.

Through the data reduction process outlined here, a global understanding of gene expression in human tissues is emerging. The whole body was crudely segmented into about 50 compartments based on organs/tissues represented in the dbEST libraries, and the distribution of gene expression was analyzed across those compartments. Using this algorithm, we found that most genes are not specifically expressed in any particular tissue/organ. This finding indicates that there are many common functions that are necessary for the continued existence of most or even all cell types, and the specialized functions that differentiate cell types are carried out by a relatively small subset of genes. Almost every tissue/organ represented in this segmentation had a few specific genes, with liver having the most. To complete the global picture of gene expression it would be important to segment the human organ/tissues into even smaller and more pure sub-segments with the help of laser capture microdissection. We expect that the distribution of gene expression will then be more accurately reflected by expression profiling methods.

ACKNOWLEDGEMENTS

This research was supported by a generous gift from The Richard M. Schulze Family Foundation. Funding for this work was also provided by the Mayo Clinic Comprehensive Cancer Center, the Department of Laboratory Medicine and Pathology. Conflict of Interest: none declared.

REFERENCES


