

Data Mining for Biomarker Development: A Review of Tissue Specificity Analysis

Eric W. Klee, PhD

*Division of Experimental Pathology, Department of Laboratory Medicine and Pathology,
Mayo Clinic, 200 1st Street SW, Stabile 2-50, Rochester, MN 55905, USA*

Tissue-specific expression profiling is the simultaneous measurement of the expression of thousands of genes in a target tissue or organ, and it can play an important role in biomarker development and in filling clinicians' unmet needs for improved marker-based assays. The National Cancer Institute's report on 2008 research, *The Nation's Investment in Cancer Research*, states that identifying candidate biomarkers that successfully translate into improved diagnostic and prognostic assay is a priority objective [1]. Such biomarkers can impact patient care and outcome through disease screening, early detection, risk stratification, and prediction of disease reoccurrence. Additionally, with the increasing trend toward individualized medicine and personalized treatment, "theragnostic" biomarkers are needed to identify patients responsive to specific therapies and to track the therapeutic effects [2,3]. To discover and bring to clinic these novel assays requires a data-driven development cycle that incorporates data-mining steps for discovery, qualification, verification, and validation [4].

Biomarker development has leveraged the enormous data content obtainable from high-throughput genomic and proteomic technologies to identify novel candidates, effectively transforming this research into a practical exercise in data mining. Genomic discovery using microarrays and proteomic discovery using mass spectrometry are becoming the primary methods for initial candidate marker selection [5–7]. The identification of differentially expressed candidate markers using these techniques is proving to be only the first step in the biomarker development process, however. Further marker characterization and prioritization are needed to increase the chances of these candidates being validated and translated into clinical assays.

E-mail address: klee.eric@mayo.edu

The data mining of public transcriptomic databases to define tissue-specific expression profiles of biomarkers provides one way to do this. This article reviews four databases that were selected because they contain different types of transcriptomic data and provide interfaces for profiling tissue-specific gene expression. This review is followed by a discussion of eight tissue-specificity studies, with detailed descriptions of the data-analysis methods and metrics used to identify and quantify tissue specificity. The article concludes with a short summary of biomarker development projects that have mined the transcriptomic databases and applied tissue-specificity metrics is provided.

Public transcriptomic databases

Transcriptomics is the study of the complete set of mRNA transcripts expressed in a tissue or cell type at one time. High-throughput sequencing methods have facilitated transcriptomic studies by generating an abundance of data describing the expression of gene transcripts within various human tissues and disease states. These data are maintained in large, public transcriptomic databases that are available to researchers for analysis. In this section, four representative databases are described to familiarize readers with the data content and interfaces available. The chosen databases each house data generated by a different expression-profiling technology, including Expressed Sequence Tag (EST) [8], Serial Analysis of Gene Expression (SAGE) [9], Massively Parallel Signature Sequencing (MPSS) [10,11], and microarray expression analysis. Although other public databases contain these data types, these databases were selected because they provide direct and clear tools for generating tissue-specific gene-expression profiles.

Unigene

Unigene is a National Institutes of Health (NIH)–sponsored database of assembled EST transcripts [12]. The database is designed to provide an organized collection of dbEST [13] sequences, processed and clustered to be associated with identifiable genes and presented in a format useful to researchers. ESTs are partially sequenced cDNAs obtained by reverse transcription of mRNA and provide a snapshot of gene expression at the time of cDNA synthesis. ESTs generally are sequenced using a single-pass approach, are 3′-biased, contain inaccuracies (approximately 2%), and are highly redundant [8]. Unigene predates the fully sequenced human genome, and previously it provided one of the most comprehensive data sets for gene discovery research. It continues to provide a compendium of information on gene expression and expression characteristics within tissue, disease states, and developmental stages, by estimating gene-expression levels using EST clone frequency in different libraries [14]. A systematic assembly method is used to assign short EST fragments to consensus sequence clusters that

are anchored at the 3' end of a gene. The present build of the database (#201, updated 03/01/2007) contains sequence data for 77 species, including more than 6,694,833 human transcripts in 124,179 clusters.

The Unigene interface for tissue-specificity analysis is well suited for single-gene queries. Each EST cluster entry provides a Gene Expression Summary section that includes links to the cluster Expression Profile. Within the profile, a numeric and graphic display of the cluster expression in different tissue libraries is provided. Three distinct profiles are provided, illustrating the relative cluster expression in different normal tissues, in various health states, and in independent developmental stages. When more than half of a cluster's entire EST counts are assigned to a single profile state, that state is labeled as a "restricted expression" state. For example, PSA (KLK3), a well-known prostate cancer serum biomarker, shows restricted expression in prostate, prostate cancer, and adult developmental stage. Unigene also can be queried for specific tissue-expression profiles using the Digital Differential Display (DDD) tool, which generates information on tissue-specific distribution of expression defined by the EST library of origin. DDD uses the Fisher Exact Test to compare the distribution of transcripts in two different libraries, thereby allowing users to compare expression in two libraries (tissue types) of interest. Analysis is restricted to "deeply sequenced" libraries (> 1000 sequences) to ensure validity of results. Overall, Unigene provides a rich information resource for biomarker data mining, generating gene-expression profiles across tissue, disease, and development stage strata.

The Cancer Genome Anatomy Project

SAGE is a transcriptional profiling technique that surveys short 3' oligonucleotides derived from mRNA to quantify gene expression. High-throughput, highly redundant sequencing generates millions of transcripts that provide information on gene expression that is complementary to traditional EST methods. The Cancer Genome Anatomy Project (CGAP) of the NIH maintains a large database of SAGE sequences [15]. An internal analysis pipeline filters out erroneous data and maps SAGE transcripts to gene sequences. Based on the relative number of transcripts mapped to a gene and the disease or tissue characteristics of the originating cellular library, tissue-specific and disease-specific expression profiles can be generated.

Within CGAP, the SAGE Genie is used to mine the database and generate several tissue-specific expression profiles. The SAGE Anatomical Viewer creates a list of "best" SAGE tags for a specific gene. For any tag, three different visualizations of the data can be selected. The Ludwig Transcript Viewer provides a positional map of SAGE transcripts on the target gene, the Digital Northern provides a frequency-sorted list of tag occurrence in SAGE libraries, and the Anatomical Viewer generates three heat-map

pictographs (tissues, cell-lines, and both), displaying the relative expression level of a gene in 24 normal and cancer tissues. Additionally, the SAGE Genie toolset includes the Differential Gene Expression Displayer and Experimental Viewer, which compare the relative expression of tags between two user-defined libraries, in the same manner as the Unigene DDD tool.

The CGAP database continues to grow as SAGE transcripts from recent research projects are added to it. For example, a study on glioblastomas generated 116,259 new SAGE tags that were summarily deposited in CGAP [16]. The study investigators then used the SAGE Genie tool suite to evaluate their expression data and identify several genes highly specific to this disease, demonstrating the power of the CGAP database.

The Ludwig Institute for Cancer Research—Massively Parallel Signature Sequencing data

MPSS is a technique based on a novel cloning and sequencing method employing enzymatic digestion and hybridization of short polyadenylated transcripts using microbeads [10]. This method is advantageous because it has a large dynamic range (10^5 –fold differences) and consequently can generate a richer gene-expression data set. The Ludwig Institute for Cancer Research (New York, New York) developed an extensive MPSS database with expression data on 32 normal human tissue types [17]. The database is Web accessible and can be queried for gene-expression profiles across all tissue types for individual genes and gene lists, as well as for expression in restricted tissue types. This database provides a valuable resource for assessing the nominal expression characteristics of a gene but does not provide any information on expression characteristics in cancer/disease states.

SymAtlas V1.2.4

The SymAtlas is a microarray-expression database supported by the Genomics Institute of the Novartis Research Foundation (Cambridge, Massachusetts) (GNF) [18]. Custom human microarray chips, designed and fabricated by Affymetrix (Santa Clara, California), were used to measure gene expression in 46 normal human tissues. All samples were prepared and analyzed using a standard protocol to enable accurate comparisons between experiments. The database can be searched in a targeted manner for single genes, and a graphic display of expression across tissue types is returned, along with a set of textual annotations including basic gene characteristics, gene ontology–linked functional descriptors, associated transcripts and proteins, a textual summary of the gene, a list of corresponding probe sets, and a list of relevant literature citations. The graphic display gives a relative measure of expression for the target gene in all tissues. In addition to single-gene queries, users can also do batch queries for general expression profiles. The batch query form allows users to define unique expression thresholds (fold change relative to median) for all tissue types, combined

by a logical AND operator. Unlike the three previously described databases, in which data are derived from short oligo-sequencing technologies, GeneAtlas is based on expression information obtained from customized hybridization arrays created by Affymetrix.

Tissue-specific expression-profiling methods

A number of studies have developed methods for analyzing the previously described data sets and generating rigorous metrics of tissue specificity. In this section, eight methods are discussed in detail (Table 1). Each method uses a unique statistical test or set of decision criteria for identifying genes with tissue-specific expression profiles.

Selective expression

The selective-expression approach is based on an algorithm for identifying genes with selective expression, defined as significantly different expression in one tissue compared with other tissues, based on four basic properties: intensity, source, source-set, and source-confidence [19]. The methods used are designed to be adaptable to different data types, including microarray data and EST, SAGE, or MPSS library transcript counts. The intensity parameter denotes the expression value for a given gene in a tissue type denoted by the source parameter. The source-set parameter allows sources to be aggregated into subsets of interest. The source-confidence parameter allows the inclusion of a quantifiable metric for weighting the quality of a measurement in the specificity analysis. For example, it can be used to correct for differences in cDNA library sequencing depth (sampling size). Statistical significance of selective expression is measured

Table 1
Methods for computing tissue specificity metrics from large gene-expression data sets

Study	Data source	Specificity metric/test
Selective expression	EST, SAGE, MPSS	Dixon discordance test
TissueInfo	EST	Expression ratio and rule-based decision tree
ExQuest	EST	TSU
GEPIS	EST	DEU - Z-statistic
Shannon Entropy	EST, microarray	Shannon Entropy and Q-statistic
Akaike's Information Criterion	Microarray	Akaike's Information Criterion
Tissue selectivity	Microarray	Honest Significant Difference
ROKU	Microarray	Shannon Entropy and Akaike's Information Criterion

Abbreviations: DEU, digital expression units; EST, Expressed Sequence Tag; GEPIS, Gene Expression Profiling in silico; SAGE, Serial Analysis of Gene Expression; MPSS, Massively Parallel Signature Sequencing; TUS, tissue-specific units.

using the Dixon discordance test for uniform distributions. The data analysis includes the following steps:

1. Filter the data points for high-quality measurements.
2. Verify a minimum number of data points passing step one for each source.
3. Apply the quantitative test of discordance and use the statistical significance score to determine if scores are reliable.
4. Adjust scores computed in step three for the baseline expression level of nonvariant genes.
5. Compute the minimum intensity gap (the separation between the largest intensity and the second largest intensity).
6. Compute an overall confidence for the selective expression parameter by combining steps four and five to provide an overall confidence level for the selectivity defined by the numeric evaluation of parameters in a single decision function.

Step four of the process allows the data to be adjusted to reduce the significance of discordant values as the baseline expression level approaches saturation for the measurement system used. Step five introduces a level of robustness to the discordant values in step six by noting if the minimum intensity gap approaches the resolution power of the measurement system used.

The decision metric described in this approach is designed to be universally applicable to any data source capable of reporting tissue-specific expression profiles. The six-step analysis method computes a numeric value for tissue “selectivity” that can be configured to the technique used to generate the data to prevent erroneous predictions. The algorithm has been used in several studies to identify tissue-selective expression [20–24].

TissueInfo

The TissueInfo method and interface were constructed to provide tissue-specific expression profiles for a target gene sequence on the basis of EST sequence counts [25]. TissueInfo uses the Basic Alignment Search Tool (BLAST) [26] and MegaBLAST [27] sequence comparison programs to align individual EST sequences obtained from dbEST to a user-submitted query sequence. Based on the number of ESTs matching the query sequence and the associated annotation of the matching EST source libraries, the query sequence is annotated as (1) expressed in a tissue, (2) specifically expressed in a tissue, or (3) tissue specific. To be labeled as expressed in a tissue, the query sequence need only have a single matching EST sequence derived from a library constructed from that tissue type. To be specific to a tissue, the number of ESTs from the tissue in question matching the query sequence divided by the total number of ESTs matching the query sequence must be greater than a user-defined threshold for specificity (the authors use

0.95). To be tissue-specific, the query sequence can be expressed only in a maximum of two tissues, and the ratio of ESTs matching the query sequence coming from a pooled tissue set over all the ESTs matching the query sequence may not exceed one minus the threshold for specificity. This method for determining tissue specificity is a straightforward and intuitive ratio test but fails to incorporate statistical tests of significance included in several of the other methods described.

A unique characteristic of TissueInfo is it computes tissue-specific expression values for both gene- and protein-query sequences by changing the sequence comparison algorithm used. The authors leveraged this characteristic to test the tissue-specificity predictions for a set of protein sequences obtained from the SwissProt database (<http://ca.expasy.org/sprot/>) with annotated Tissue Specificity. This testing showed the program was 69% accurate in predicting whether a query sequence was specific to (expressed predominantly in) a tissue and 80% accurate in predicting whether a query sequence was expressed in a tissue [25]. It is possible the accuracy estimates are lower than the true program accuracy, because it could be questioned whether the Tissue Specificity annotations for the proteins evaluated are based on sufficient experimental depth or tissue source breadth to be properly comparable with the range of tissues sampled by EST sequences and used in the TissueInfo calculations.

TissueInfo is hosted by the Weill Medical College of Cornell University (<http://icb.med.cornell.edu/crt/tissueinfo/webservice.xml>) and has been maintained with current updates to the underlying sequence sets. The interface allows users to generate a tissue-expression profile for a specific sequence or search the database for genes that are expressed or predominantly expressed in designated tissues.

ExQuest

The ExQuest program maps ESTs to a target sequence and uses the underlying EST library annotations to compute tissue-specificity profiles [28]. Sequence mapping is performed by MegaBLAST. Based on these comparisons, the actual number of ESTs matching the query sequence for each tissue category is recorded. These values subsequently are normalized for the relative depth of the underlying EST library sampling, using an Expected Hit metric. This metric is calculated for each tissue category by multiplying the ratio of the number of ESTs from a given tissue type over the total number of ESTs included in the ExQuest database by the total number of ESTs matching the query sequence. The result is an estimated EST hit rate for each tissue based completely on the underlying EST library sizes. Finally, for each tissue, the actual EST hits are divided by the expected EST hits to obtain a normalized metric of expression called tissue-specific units (TSU). Based on the distribution of TSU values across tissues, users can identify tissue-specific expression patterns for the query sequence.

ExQuest does not incorporate any statistical test for tissue specificity. The TSU metric provides an interesting and simple approach to normalizing expression measurements between libraries that are sampled to different extents. It should provide a fair estimate of tissue-specific expression levels, provided the analysis includes no libraries that are sampled at a very low level.

ExQuest provides a unique hierarchical organization of tissue categories embedded in the program. EST libraries are organized into hierarchical tissue bins in which related libraries are grouped together, thereby increasing the overall transcript count per tissue category. The hierarchical organization is designed to provide users with three levels of organization for querying specificity: at the primary tissue, the secondary tissue, and individual EST library. For example, a primary tissue would be the pancreas, a secondary tissue would be the islet of Langerhans, and the individual EST library would be a single EST library derived from the islet of Langerhans. This hierarchical organization gives users the flexibility to balance sampling depth in a category with more exact feature specificity. A second unique feature of ExQuest is that users can define the degree of similarity by which EST sequences are mapped to gene entities by selecting the percentage alignment parameter used by MegaBLAST. The authors demonstrate that varying this parameter allows related paralog sequences to be differentiated from each other. Finally, ExQuest also provides an interesting interface option to view chromosomal mapping of sequence-associated ESTs. This feature enables users to zoom in and out of a chromosomal map while selecting specific tissues for which matching ESTs will be displayed. This interface lets users evaluate tissue-specific expression patterns in a positional context and associate these patterns with established genomic effects. The ExQuest program is available through a Web server (<http://lena.jax.org/~dcb/ensRNA/exquest.html>) but has not been updated since 2004 (Derry Roope-nian, PhD, personal communication, Bar Harbor, Maine, 2007). The hierarchical grouping of libraries is a unique and useful feature not explicitly provided in other tissue-specificity methods, allowing users to adapt the analysis to the level best suited to their biomarker study. It is unfortunate the system is not in active development, but even in its present state it may be a useful resource for investigators.

Gene Expression Profiling in silico

The Gene Expression Profiling in silico (GEPIS) method uses EST sequence mapping to compute tissue-specific expression in 43 tissue types [29]. Before computing specificity, the GEPIS system executes a rigorous data-quality filtering protocol on the ESTs and EST libraries obtained from dbEST. All ESTs obtained from libraries with tissue-source annotations of unknown, ambiguous, or pooled tissue type are excluded from the analysis. EST libraries that have been normalized or “subtracted” and

EST libraries derived from fetal or embryonic tissues also are excluded from the analysis. Finally, the authors also eliminated several EST libraries that were deemed to have been misannotated according to their expression analysis. The remaining EST sequences were assigned to gene sequences using BLAST comparisons.

GEPIS computes EST library normalized expression values, called “digital expression units” (DEU), for all genes in all tissues. The DEU is equal to the number of ESTs assigned to a gene in a tissue, divided by the total ESTs in the tissue, and multiplied by a scaling factor of 1,000,000. A Z-statistic then is used to compare any two tissue categories and determine statistical significance. The Z-statistic is computed by Equation 1, and the result compared with a normal distribution to obtain a p-value.

$$Z = \frac{\hat{p}_A - \hat{p}_B}{\sqrt{\hat{p}(1 - \hat{p})\left(\frac{1}{N_A} + \frac{1}{N_B}\right)}} \quad (1)$$

Where $\hat{p}_A$ and $\hat{p}_B$ are the DEU values for tissues A and B, $\hat{p}$ is the DEU for the gene of interest across all tissues, and N_A and N_B are total number of EST sequences from tissues A and B. For a given gene, the Web interface returns a graphic representation of the tissue-specific expression profile. It also provides a chart containing raw EST counts and DEU values for both normal and cancer tissue libraries, a p-value based on the Z-statistic comparing the normal and cancer DEU values within a tissue, and the highest p-value obtained from the Z-statistics comparing the gene expression in a normal tissue with the expression in each of the other normal tissues.

GEPIS, much like the ExQuest program [28], computes a “Regional Atlas” that graphically shows the expression of all genes in proximity to the gene of interest, across a user-defined range and number of tissue types. This feature is useful when investigating events such as cancer-induced copy-number effects. The program authors also attempted to verify GEPIS gene-expression profiles experimentally using quantitative polymerase chain reaction (qPCR) measurements of 40 genes over a range of expression levels in normal and cancer colon samples [29]. The analysis revealed strong correlation between the GEPIS expression level and qPCR expression level at all expression ranges (low, medium, and high). GEPIS recently has been revised and released as GeneHub-GEPIS [30]. The upgraded version now allows users to search the database using multiple database identifiers.

Shannon Entropy

Schug and colleagues used Shannon Entropy (generally designated H in information theory) was used to measure the overall specificity of a gene, that is, the degree by which a gene-expression profile differs from a ubiquitous expression profile [31]. This statistical measure of specificity provides

a single metric for assessing the complete gene-expression profile but does not provide any information regarding the tissues in which a gene may be specifically expressed. To obtain this information, a new statistic, Q , is introduced to measure what the authors label “categorical specificity” or expression specific to a tissue type. This method was used to measure tissue specificity in both the GNF Gene Expression Atlas [18] microarray data and the Database of Transcribed Sequences, EST data [32]. The microarray expression values were analyzed, without modification, as obtained from the Gene Expression Atlas. For the EST data, counts of EST sequences associated with a given gene in a given tissue were normalized into pseudo-counts by Equation 2.

$$w_{g,t} = \frac{(n_{g,t} + 1)}{(N_t + N_g)} \quad (2)$$

Where $n_{g,t}$ is the number of EST sequences mapped to gene g , in tissue t . N_t is the total number of EST sequences associated with tissue t , and N_g is the total number of genes. Relative expression values $p_{t|g}$ were computed for all expression values by dividing $w_{g,t}$ by the total expression for a gene across all tissues. Using the relative expression values, the Shannon Entropy was computed using Equation 3.

$$H_g = \sum_{1 \leq t \leq N} -p_{t|g} \log_2(p_{t|g}) \quad (3)$$

Where N is the total number of tissues evaluated. A low entropy value reflects a highly specific gene-expression profile. The Q -statistic, for measuring categorical specificity, then is computed from the Shannon Entropy using Equation 4.

$$Q_{g|t} = H_g - \log_2(p_{t|g}) \quad (4)$$

Where a Q -value of zero denotes expression restricted to that tissue only, and an increasing Q -value is indicative of a more ubiquitous expression profile.

A comparison of specificity measurements between the microarray and EST data using the same gene set and tissue set showed the overall distribution of entropy across all genes was slightly depressed in the EST data set versus the microarray data set. This finding may reflect insufficient sampling associated with some EST sequence libraries, so that several genes seem to be expressed in a more specific manner than found when analyzing the microarray data set. Independent experimental measurements would be required to understand the discrepancies between the specificity measurements from the two data sources.

The method described in this article introduces a classic measure from information theory, Shannon Entropy, to detect tissue-specific gene-expression patterns and a novel statistic, Q , to measure tissue-specific expression.

By using these two statistics together, genes that are expressed in a tissue-selective manner can be identified and ranked by the overall measure of specificity, in a manner similar to that developed to analyze the Gene Logic (<http://www.genelogic.com>) microarray data set [33]. It would be interesting to apply the Shannon Entropy method to the GeneLogic data and compare the overall ranks and measures of gene specificity computed by the two different approaches. The method also is advantageous, because it is applicable to data from two different sources (EST and microarray).

Akaike's Information Criterion

Akaike's Information Criterion can be used to identify genes with markedly different expression in one or a few genes relative to the gene expression in most tissue types [34]. The criterion originally was developed to identify an optimal model from a class of competing models [35] but has been adapted to the detection of outlier gene expression. The authors rationalize the use of this criterion because it operates without requiring a significance level to be selected, thereby providing an objective method of gene selection. In a study of mouse microarray data obtained from the Riken Expression Array Database [36], 49 samples from different normal tissues were evaluated with the criterion. After filtering out low-quality data and normalizing to a reference array, genes of interest were selected using a U-statistic, where a low score was desirable, as defined in Equation 5.

$$U = n \log \sigma + s \sqrt{2} \left(\frac{\log n!}{n} \right) \quad (5)$$

Where $(n+s)$ equals the total number of observations across tissue types, s equals the number of outlier candidates, and σ equals the SD across scores from the n tissues (but not the s tissues)

A U-statistic was computed for up to $X(X+1)/2$ combinations of possible outliers, where $X = 1+(n+s)/2$. X varied according to the number of data points (ie, changed if there were missing data points) and used a combination of outliers that were both up- and down-regulated. This method can be adapted easily to the analysis of any large data set (eg, microarray data) for which expression ratios in multiple tissues are available or computable. The method is fairly straightforward and easy for other users to implement, and the authors believe that the lack of a p-value for specificity is an advantage. It does, however, require a significant amount of computation, up to $N*X(X + 1)$ calculations where X is the number of tissues types evaluated, and N is the number of genes evaluated.

Tissue selectivity

A robust method to calculate tissue selectivity uses the GeneLogic microarray expression data set [33]. The methods were designed to identify

tissue-selective gene expression. Tissue-selective expression is defined as modified gene expression in one or a few biologically similar tissue types and is distinguished from tissue-specific expression, defined as gene expression restricted to one and only one tissue type. The Tukey-Kramer's honest significant difference (HSD) test is used to compute pairwise comparisons of a gene's expression between all tissue types. The HSD test generates an adjusted p-value and Q-parameters for each comparison; the Q-parameter is equivalent to the absolute difference of two group means over the sample size-adjusted SD. Tissue-selective genes were identified subsequently using two rules based on the HSD p-values. First, 91 of 97 ($N = 98$) pairwise comparisons were required to have a p-value below the user-defined significance level. Second, the first rule could not be true for a gene in more than six different tissue types. Additionally, the authors used the HSD Q-scores to compute an overall probability value, called the "enrichment score," for a gene, to provide a measure of error control. This calculation was computed as shown in Equation 6.

$$ES = \frac{1}{N-1} \sum_{i=1}^N \left(1 - \frac{Q_i - \text{Min}(Q_i)}{\text{Max}(Q_i) - \text{Min}(Q_i)} \right) \quad (6)$$

Where $N = k(k-1)/2$, and k equals the number of tissues evaluated. This method for assessing tissue-specific expression patterns is advantageous for two main reasons. The authors employed an expression-comparison test that includes statistical rigor to reduce erroneous predictions. The authors also defined a method of tissue-specific expression analysis that is flexible enough to allow a gene to be expressed specifically in a few tissues, not in just a single tissue. Additionally, users can adapt the selection criteria easily to increase the stringency or flexibility in defining exactly what tissue selective expression is.

ROKU

ROKU is a method for analyzing microarray data to detect tissue-specific gene expression that combines the Shannon Entropy method [31] for ranking genes with specific expression patterns and the Akaike's Information Criterion method [34] for detecting specific tissues with expression outliers [37]. The authors rationalize the approach by claiming that the Q-statistic developed to complement the Shannon Entropy fails to acknowledge redundancy in the selection of genes specifically expressed in a tissue, and this redundancy can be handled better with an outlier selection criterion. Furthermore, combining these two techniques into a single method enables the detection of tissue-specific expression that is over- or underexpressed relative to the nominal gene expression. Consequently, the ROKU method is designed not to detect genes that are expressed in a small number of tissues and are not expressed in the remaining tissue but to detect genes that are

differentially expressed in a small number of tissues relative to a generally stable expression in most tissues. That is, the ROKU method is designed to identify genes expressed at significantly higher or significantly lower levels in a small number of tissues relative to the general background expression level of that gene, a process the authors call "broad sense specificity" [37].

The tissue-specificity calculations use the same equation for Shannon Entropy (H) as previously described in Equation 4. The only difference is that the ROKU method first transforms the microarray expression values, $w_{g,t}$, by subtracting the one-step Tukey biweight and taking the absolute value. This processing effectively recenters the data and enables the Shannon Entropy metric to detect genes that are specifically expressed in a broad sense. The tissues in which genes are specifically expressed then are identified using the Akaike's Information Criterion for outlier detection as described in Equation 1.

A clear disadvantage in the ROKU method is that the one-step Tukey biweight transformation prevents the analysis from identifying binomial expression patterns, in which a gene is expressed at one level in half the tissues and at a different level in the remaining tissues. The transformation adjusts the expression vector, so that the overall expression profile appears uniform. The ROKU method, however, has other advantages over the Shannon Entropy method and the Akaike's Information Criterion method, because it can detect tissue-specific gene expression and rank genes according to overall tissue specificity and also can identify genes that are specifically expressed in only one tissue and not simply in a small number of tissues.

Summary

The studies described in this article provide an introductory review of data-mining techniques that have been applied to large transcriptomic data sets to derive tissue-specific expression profiles. The methods were selected to illustrate different decision criteria and data preprocessing applied to both microarray and short oligonucleotide (EST, SAGE, MPSS) data sets. Several other studies on tissue-specificity analysis have applied a spectrum of methods to the problem, from simple fold-change [38] and t -test [39] decision criteria to more complex principal component analysis [40] and binary indexing [41] algorithms. In addition, several programs have been developed to improve user access and provide more intuitive tissue-profiling functionality to the large transcriptomic databases previously described [42–44]. These methods provide researchers with a range of gene-expression data and analysis tools that can drive or complement biomarker development projects.

Tissue-specific expression profiling has been discussed using four different types of data. Most analysis methods were developed to analyze a specific data type, although several demonstrated interoperability between different data types. When undertaking a biomarker development study, however,

the question remains as to which data type should be analyzed to benefit the study best. Clearly, this decision depends on the specific nature of the study and whether the study-specific tissue is sampled—and sampled sufficiently—in a given database. In some cases, the study target could fall within a very specific tissue type and may be restricted to a single data source, such as the EST database or the CGAP SAGE collection. In most cases, however, any one of the four data types probably will be able to provide pertinent information. To address whether this information is redundant or complementary, Huminiecki and Bicknell [45] evaluated the congruence of specificity data obtained from SAGE and EST data analysis and later compared these data with microarray data analysis. The initial study concluded that evaluating SAGE and EST data together provided a more accurate assessment of tissue specificity than could be obtained from either data set alone. In a later study, the correlation of specificity analysis by EST and SAGE data compared with that of microarray data was strong in tissues that were extensively sampled [46]. The study reported, however, that correlation was not strong between microarray and the EST and SAGE data types in tissues that had complex cellular composition or that had not been sampled extensively. This finding suggests that EST and SAGE libraries measure up to microarray analysis only in tissues that are deeply sampled within those libraries. None of the microarray analysis methods discussed here have addressed the possible effect of low-level microarray measurements on the robustness of the tissue-specificity predictions. Low-expression measurements often fall within the background level of an array technology, and any analysis dependent on these signals is potentially erroneous. Consequently, it would seem that for genes expressed at a low level or in a poorly sampled tissue, analysis of a single data type would be insufficient. Therefore, researchers should consider evaluating tissue-specific expression in multiple (or all available) data types to obtain the most comprehensive expression profile.

Tissue-specificity expression profiling has been used widely for biomarker discovery but is equally or more applicable to candidate biomarker characterization. The literature is filled with references to studies that have identified candidate biomarkers by mining transcriptomic data for genes differentially expressed in normal and cancerous tissue [47–55]. This type of candidate biomarker discovery can be undertaken using the data-mining methods described herein. It is the purpose of this review, however, to encourage investigators to consider using these data-mining methods to generate tissue-specific expression profiles that can complement existing efforts to discover candidate biomarkers. This approach has been used to identify candidate cardiac markers that are specifically expressed in the heart [55], to identify candidate prostate cancer biomarkers that are differentially expressed in normal and cancer tissues and also are selectively expressed in prostate [54], to identify candidate brain-injury markers specific to the brain [53], and to identify candidate bladder carcinoma biomarkers specific

to the bladder [48]. Furthermore, specificity analysis has been suggested as an important metric for identifying candidate serum cancer biomarkers [56]. In all these instances, tissue-specific expression profiling provides a method for further refining lists of candidate biomarkers and intelligently selecting an enriched set of candidates to move forward in the biomarker development cycle.

References

- [1] National Cancer Institute 2007. The nation's investment in cancer research. A plan and budget proposal for fiscal year 2008. Pub. L. No. 92–218, NIH Publication No. 06-6090.
- [2] Batchelder K, Miller P. A change in the market—investing in diagnostics. *Nat Biotechnol* 2006;24(8):922–6.
- [3] Ozdemir V, Williams-Jones B, Glatt S, et al. Shifting emphasis from pharmacogenomics to theragnostics. *Nat Biotechnol* 2006;24(8):942–6.
- [4] Rifai N, Gillette MA, Carr SA. Protein biomarker discovery and validation: the long and uncertain path to clinical utility. *Nat Biotechnol* 2006;24(8):971–83.
- [5] Cho WC. Contribution of oncoproteomics to cancer biomarker discovery. *Mol Cancer* 2007; 6(1):25.
- [6] Bharti A, Ma PC, Salgia R. Biomarker discovery in lung cancer-promises and challenges of clinical proteomics. *Mass Spectrom Rev*; 2007.
- [7] He YD. Genomic approach to biomarker identification and its recent applications. *Cancer Biomark* 2006;2(3–4):103–33.
- [8] Adams MD, Kellye JM, Gocayne JD, et al. Complementary DNA sequencing: expressed sequence tags and human genome project. *Science* 1991;252(5013):1651–6.
- [9] Velculescu VE, Zhang L, Vogelstein B, et al. Serial analysis of gene expression. *Science* 1995; 270(5235):484–7.
- [10] Brenner S, Johnson M, Bridgham J, et al. Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. *Nat Biotechnol* 2000;18(6):630–4.
- [11] Jongeneel CV, Iseli C, Stevenson BJ, et al. Comprehensive sampling of gene expression in human cell lines with massively parallel signature sequencing. *Proc Natl Acad Sci U S A* 2003;100(8):4702–5.
- [12] Pontius JU, Wagner L, Schuler GD. UniGene: a unified view of the transcriptome. In: *The NCBI handbook*. Bethesda (MD): National Center for Biotechnology Information; 2003.
- [13] Boguski MS, Lowe TM, Tolstoshev CM. dbEST—database for “expressed sequence tags”. *Nat Genet* 1993;4(4):332–3.
- [14] Adams MD, Kerlavage AR, Fields C, et al. 3,400 new expressed sequence tags identify diversity of transcripts in human brain. *Nat Genet* 1993;4(3):256–67.
- [15] Boon K, Osorio EC, Greenhut SF, et al. An anatomy of normal and malignant gene expression. *Proc Natl Acad Sci U S A* 2002;99(17):11287–92.
- [16] Beaty RM, Edwards JB, Boon K, et al. PLXDC1 (TEM7) is identified in a genome-wide expression screen of glioblastoma endothelium. *J Neurooncol* 2007;81(3):241–8.
- [17] Jongeneel CV, Delorenzi M, Iseli C, et al. An atlas of human gene expression from Massively Parallel Signature Sequencing (MPSS). *Genome Res* 2005;15(7):1007–14.
- [18] Su AI, Cooke MP, Ching KA, et al. Large-scale analysis of the human and mouse transcriptomes. *Proc Natl Acad Sci U S A* 2002;99(7):4465–70.
- [19] Greller LD, Tobin FL. Detecting selective expression of genes and proteins. *Genome Res* 1999;9(3):282–96.
- [20] Stekel DJ, Git Y, Falciani F. The comparison of gene expression from multiple cDNA libraries. *Genome Res* 2000;10(12):2055–61.
- [21] Castensson A, Emilsson L, Preece P, et al. High-resolution quantification of specific mRNA levels in human brain autopsies and biopsies. *Genome Res* 2000;10(8):1219–29.

- [22] Lai C, Chou C, Ch'ang L, et al. Identification of novel human genes evolutionarily conserved in *Caenorhabditis elegans* by comparative proteomics. *Genome Res* 2000;10(5):703–13.
- [23] Walker MG, Volkmut W, Sprinzak E, et al. Prediction of gene function by genome-scale expression analysis: prostate cancer-associated genes. *Genome Res* 1999;9(12):1198–203.
- [24] Ewing RM, Kahla AB, Poirot O, et al. Large-scale statistical analyses of rice ESTs reveal correlated patterns of gene expression. *Genome Res* 1999;9(10):950–9.
- [25] Skrabanek L, Campagne F. TissueInfo: high-throughput identification of tissue expression profiles and specificity. *Nucleic Acids Res* 2001;29(21):E102.
- [26] Altschul SF, Gish W, Miller W, et al. Basic local alignment search tool. *J Mol Biol* 1990;215:403–10.
- [27] Zhang Z, Schwartz S, Wagner L, et al. A greedy algorithm for aligning DNA sequences. *J Comput Biol* 2000;7:203–14.
- [28] Brown AC, Kai K, May ME, et al. ExQuest, a novel method for displaying quantitative gene expression from ESTs. *Genomics* 2004;83(3):528–39.
- [29] Zhang Y, Eberhard DA, Frantz GD, et al. GEPIS—quantitative gene expression profiling in normal and cancer tissues. *Bioinformatics* 2004;20(15):2390–8.
- [30] Zhang Y, Luoh SM, Hon L, et al. GeneHub-GEPIS: digital expression profiling for normal and cancer tissues based on an integrated gene database. *NAR* 2007;35:W152–8.
- [31] Schug J, Schuller WP, Kappen C, et al. Promoter features related to tissue specificity as measured by Shannon Entropy. *Genome Biol* 2005;6(4):R33.
- [32] The Computational Biology and Informatics Laboratory. AllGenes: a Web site providing access to an integrated database of known and predicted human (release 9.0, 2004) and mouse genes (release 10.0, 2004). Center for Bioinformatics, University of Pennsylvania. Available at: <http://www.allgenes.org>. Accessed November 19, 2007.
- [33] Liang S, Li Y, Be X, et al. Detecting and profiling tissue-selective genes. *Physiol Genomics* 2006;26(2):158–62.
- [34] Kadota K, Nishimura S, Bono H, et al. Detection of genes with tissue-specific expression patterns using Akaike's information criterion procedure. *Physiol Genomics* 2003;12(3):251–9.
- [35] Akaike H. Information theory and an extension of the maximum likelihood principle. *Proc: 2nd Int symp information theory*. Budapest; 1973. p. 267–81.
- [36] Miki R, Kadota K, Bono H, et al. Delineating developmental and metabolic pathways in vivo by expression profiling using the RIKEN set of 18,816 full-length enriched mouse cDNA arrays. *Proc Natl Acad Sci U S A* 2001;98(5):2199–204.
- [37] Kadota K, Ye J, Nakai Y, et al. ROKU: a novel method for identification of tissue-specific genes. *BMC Bioinformatics* 2006;7:294.
- [38] Saito-Hisaminato A, Katagiri T, Kakiuchi S, et al. Genome-wide profiling of gene expression in 29 normal human tissues with a cDNA microarray. *DNA Res* 2002;9(2):35–45.
- [39] Hsiao LL, Dangond F, Yoshida T, et al. A compendium of gene expression in normal human tissues. *Physiol Genomics* 2001;7(2):95–6.
- [40] Misra J, Schmitt W, Hwang D, et al. Interactive exploration of microarray gene expression patterns in a reduced dimensional space. *Genome Res* 2002;12(7):1112–20.
- [41] Vasmatzis G, Klee E, Kube DM, et al. Quantitating tissue specificity of human genes to facilitate biomarker discovery. *Bioinformatics* 2007;23(11):1348–55.
- [42] Gupta S, Vingron M, Haas SA. T-STAG: resource and Web-interface for tissue-specific transcripts and genes. *Nucleic Acids Res* 2005;33(Web Server issue):W654–8.
- [43] Wang J, Liang P. DigiNorthern, digital expression analysis of query genes based on ESTs. *Bioinformatics* 2003;19(5):653–4.
- [44] Madden SF, O'Donovan B, Furney SJ, et al. Digital extractor: analysis of digital differential display output. *Bioinformatics* 2003;19(12):1594–5.
- [45] Huminiecki L, Bicknell R. In silico cloning of novel endothelial-specific genes. *Genome Res* 2000;10(11):1796–806.

- [46] Huminiecki L, Lloyd AT, Wolfe KH. Congruence of tissue expression profiles from gene expression atlas, SAGEmap and TissueInfo databases. *BMC Genomics* 2003;4(1):31.
- [47] Campagne F, Skrabanek L. Mining expressed sequence tags identifies cancer markers of clinical interest. *BMC Bioinformatics* 2006;7:481.
- [48] Wang XS, Zhang Z, Wang HC, et al. Rapid identification of UCA1 as a very sensitive and specific unique marker for human bladder carcinoma. *Clin Cancer Res* 2006;12(16):4851–8.
- [49] Wang AG, Yoon SY, Oh JH, et al. Identification of intrahepatic cholangiocarcinoma related genes by comparison with normal liver tissues using expressed sequence tags. *Biochem Biophys Res Commun* 2006;345(3):1022–32.
- [50] Yoon SY, Kim JM, Oh JH, et al. Gene expression profiling of human HBV- and/or HCV-associated hepatocellular carcinoma cells using expressed sequence tags. *Int J Oncol* 2006; 29(2):315–27.
- [51] Huang ZG, Ran ZH, Lu W, et al. Analysis of gene expression profile in colon cancer using the cancer genome anatomy project and RNA interference. *Chin J Dig Dis* 2006;7(2):97–102.
- [52] Aouacheria A, Navratil V, Barthelaix A, et al. Bioinformatic screening of human ESTs for differentially expressed genes in normal and tumor tissues. *BMC Genomics* 2006;7:94.
- [53] Laterza OF, Modur VR, Crimmins DL, et al. Identification of novel brain biomarkers. *Clin Chem* 2006;52(9):1713–21.
- [54] Asmann YW, Kosari F, Wang K, et al. Identification of differentially expressed genes in normal and malignant prostate by electronic profiling of expressed sequence tags. *Cancer Res* 2002;62(11):3308–14.
- [55] Megy K, Audic S, Claverie JM. Heart-specific genes revealed by expressed sequence tag (EST) sampling. *Genome Biol* 2002;3(12):RESEARCH0074:1–11.
- [56] Klee EW, Finlay J, McDonald C, et al. Bioinformatics methods for prioritizing serum bio-marker candidates. *Clin Chem* 2006;52(11):2162–4.