VeryGene: linking tissue-specific genes to diseases, drugs, and beyond for knowledge discovery

Xiaoqin Yang,1 Yun Ye,1,2 Guiping Wang,1,3 Hong Huang,4 Dekuang Yu,5 and Shuang Liang1

1Institute of Genetic Engineering, Southern Medical University, Guangzhou, Guangdong Province; 2Department of Biological and Chemical Engineering, Guangxi University of Technology, Liuzhou, Guangxi Province; 3Guangzhou Medical College, Guangzhou, Guangdong Province, People’s Republic of China; 4School of Information, University of South Florida, Tampa, Florida; and 5Southern Medical University, Guangzhou, Guangdong Province, People’s Republic of China

Submitted 9 September 2010; accepted in final form 18 January 2011

VeryGene: linking tissue-specific genes to diseases, drugs, and beyond for knowledge discovery, Physiol Genomics 43: 457–460, 2011. First published January 18, 2011; doi:10.1152/physiolgenomics.00178.2010.—In addition to many other genes, tissue-specific genes (TSGs) represent a set of genes of great importance for human physiology. However, the links among TSGs, diseases, and potential therapeutic agents are often missing, hidden, or too scattered to find. There is a need to establish a knowledgebase for researchers to share this and additional information in order to speed up discovery and clinical practice. As an initiative toward systems biology, the VeryGene web server was developed to fill this gap. A significant effort has been made to integrate TSGs from two large-scale data analyses with respective information on subcellular localization, Gene Ontology, Reactome, KEGG pathway, Mouse Genome Informatics (MGI) Mammalian Phenotype, disease association, and targeting drugs. The current release carefully selected 3,960 annotated TSGs derived from 127 normal human tissues and cell types, including 5,672 gene-disease and 2,171 drug-target relationships. In addition to being a specialized source for TSGs, VeryGene can be used as a discovery tool by generating novel inferences. Some inherently useful but hidden relations among genes, diseases, drugs, and other important aspects can be inferred to form testable hypotheses. VeryGene is available online at http://www.everygene.com.

HUMAN TISSUES exhibit distinct characteristics despite differentiating from a common origin to fulfill the different needs of our body. This kind of diversity is contributed largely by the coordinated expression of different tissue-specific genes (TSGs), in addition to other genes. The tissue-specific expression pattern of a gene implies not only its physiological function(s) but also where it plays roles in transcriptional regulation, development, stress response, and even disease etiology. Evidence gathered through mining tissue specificity, gene connectivity, and disease association suggests that many disease-associated genes are likely to show specific expression in the tissues from which the diseases originate (9, 18). Furthermore, several studies have utilized tissue specificity as an important factor when characterizing therapeutic/drug targets (5, 27). Other areas for the use of tissue specificity include, but are not limited to, pathogenic mechanism, diagnosis, and therapeutic applications (17, 21, 23).

A number of databases have been created to facilitate studies of TSGs. For example, BioGPS (24), TiGER (11), COXPRESSIONdb (15), and TiSGeD (26) databases can be used to query human gene expression in various tissues. However, most of the databases focus on the specific expression patterns of TSGs, whereas other important biological aspects are not much emphasized. For those who would like to study protein-function, protein-localization, protein-disease, or drug-target association together, the above databases alone could not serve the users’ best interest. This hinders the practical use of TSGs in medical research and the development of human systems biology. Therefore, a discovery tool dedicated to linking and providing all of the above information is highly desirable.

Here we present a web-accessible TSG knowledge discovery tool, VeryGene. It is the result of a systematic effort to integrate TSGs surveyed across a large panel of normal human tissues with other important information including subcellular localization, functional annotation, disease/drug relation, and so forth. VeryGene serves as a TSG-specific knowledgebase and a discovery tool to generate testable hypotheses for basic and clinical research.

METHODS

Although there are several TSG sets identified from other independent studies, the respective coverage of sample and tissue number is often limited. This makes it harder to conclude whether or not such TSGs are truly expressed in a tissue-selective/specific pattern. In 2004 and 2006, Su et al. (20) and our previous study (10) independently generated a tissue-specific/selective mRNA expression matrix of thousands of genes across a large panel of biological samples (~4,000 samples combined) and tissue types (~130 tissue types combined) from normal human subjects through microarray expression profiling analysis. Therefore, only these two data sets were selected for integration because of their extended coverage.

Analytically, searching for TSG amounts to comparing gene expression over many tissue types. To determine the tissue distribution for a given gene j across K tissue types, there exist P = K(K − 1)/2 pairwise comparisons for K tissue types. In our previous analyses (10), a modified Tukey-Kramer honest significant difference (HSD) test with an enrichment score (ES, Eq. 1) was proposed to overcome the type I error from multiple tests. An HSD test generates one Q value (difference of means between tissue pairs over standard deviation) per pairwise comparison for each gene. An ES [ES ≤ (0, 1)] takes into account Q values from P pairwise comparisons in one HSD test to represent tissue selectivity of a gene. The higher the value of ES for a gene, the more selective it would be. To minimize the type I error from multiple HSD tests, each and every TSG, was identified by observation of an ES greater than that by chance alone (estimated by permutation). zj scores (Eq. 2) were calculated to represent the relative level of a given TSG, expressed in one particular tissue j (j = 1 ~ K) with regard to the mean expression of TSG, across all tissues.
K tissues. The product of $z_{ij}$ and $ES_i$, denoted as $\tau_{ij}$ (Eq. 3), was computed to account for both tissue specificity/selectivity and relative expression level of a TSG in a given tissue $j$. A large $\tau_{ij}$ specifies that a TSG is highly specific and significant to a tissue $j$. In accordance with this quantitative index, not only genes specific to a tissue but also tissues in which a gene selectively expressed could be ranked (available only in Tissue View).

$$ES_i = \frac{1}{P - 1} \sum_{i=1}^{P} \left( 1 - \frac{Q_p - \text{Min}(Q_p)}{\text{Max}(Q_p) - \text{Min}(Q_p)} \right) \quad (1)$$

$$z_{ij} = \frac{x_{ij} - \mu_j}{\sigma_j} \quad (2)$$

$$\tau_{ij} = z_{ij} \times ES_i \quad (3)$$

Probe IDs were mapped to Entrez Gene IDs. Tissue names were carefully unified according to standard anatomic terms, and redundant tissue affiliations were merged according to the mean value of $\tau$. Finally, 3,960 TSGs were identified through expression profiling of a panel of 127 human tissue and cell types. These TSGs express selectively in approximately two tissues on average.

To elucidate the functional aspects of these TSGs, detailed annotations were collected. Features of each specific gene are available at six levels: subcellular localization, Gene Ontology (GO) annotation, biological pathways, mammalian phenotype linkage, disease association, and targeting drugs. Subcellular localization information for these TSGs was retrieved from LOCATE (19), supplemented with cellular component annotation of the GO database (6). Molecular function and biological process were also obtained from GO. The pathway and reaction information came from KEGG (8) and Reactome (12), respectively. Mammalian phenotype information was derived from Mouse Genome Informatics (MGI) (3). Gene-disease relationships were gathered from Gene2MeSH (1), Online Mendelian Inheritance in Man (OMIM) information (7), and Swiss-Prot (14).

Nonstandard disease names were associated with MeSH IDs and mapped to the MeSH tree categories. Gene-targeting drug relationships were obtained from DrugBank (22). With integration of these data, 5,672 gene-disease relationships and 2,171 gene-drug relationships have been collected (Table 1).

### RESULTS AND DISCUSSION

**Application.** The VeryGene server was implemented in PHP/SQL and is web accessible through an intuitive interface. The data contents were configured into two basic views, Gene View and Tissue View, to allow users to conveniently retrieve information relevant to a single gene and tissue/subcellular localization of interest, respectively. Of particular note, Batch View, which evaluates the enrichment of tissue specificity, subcellular localization, pathway, GO, phenotype, disease, and...
drug for many genes in a single query, is also provided for users to analyze genes of interest. Batch View is useful to find hidden links and to generate hypotheses. Multiple View was also developed to allow users to conduct richer combinatorial queries meeting several biological characteristics simultaneously. This can be used to look up complex relationships, and it facilitates discoveries such as “Which proteins of pathway X OR subcellular location Y are tissue-specific?” and so forth.

The resultant genes can subsequently be used to perform enrichment analysis with Batch View. Wildcard search is supported under suitable circumstances. Results from all views, as well as the entire data set used to build VeryGene, are downloadable for off-line use.

The TSGs closely related to a specific disease could have hidden links to other biomarkers or therapeutic targets/agents. VeryGene allows us to identify these unexpected links in order to generate new hypotheses. In the following example, eight TSGs for periodontitis (MeSH: D010518) could be found from Multiple View. Batch View analysis shows that five genes among them are also related to rheumatoid arthritis (MeSH: D001172). These five TSGs are enriched in such biological processes as immune response and inflammatory response. In addition, they share some common biological pathways, such as cytokine-cytokine receptor interaction and toll-like receptor signaling pathway for the two diseases. Indeed, these findings are consistent with emerging evidence of periodontitis and rheumatoid arthritis sharing many pathological features and biological links (4, 13, 25). The Batch View result also suggests that certain TNF inhibitors (e.g., etanercept and adalimumab) suitable for one medical condition might be useful for another.

For instance, recent studies showed that periodontal therapy using these inhibitors reduced the severity of active rheumatoid arthritis in patients (Fig. 1) (16). Besides, it is well known that drug development is time-consuming and very expensive. Finding new indications for existing drugs may help to capitalize the use of such drugs to remedy other medical conditions. Another example presented here is regarding simvastatin (DrugBank: DB00641), which is a hypolipidemic drug used to control hypercholesterolemia and to prevent cardiovascular disease. A sequential Multiple View/Batch View analysis indicates that 10 TSGs enriched with simvastatin also enriched with 8 most significant MeSH/MIM terms ($P < 10^{-5}$). Most of these terms can be broadly classified into vascular or inflammatory diseases, among which endometriosis (MeSH: D004715) (Fig. 2) distinguishes itself from the others as being suggested to be an autoimmune disease. The potential protective effect of simvastatin on endometriosis was preliminarily verified by study in a nude mouse model (2). These examples clearly demonstrate the power of VeryGene to reveal the hidden links some of the earlier databases failed to capture. Many questions such as “How many pathways are enriched in tissue A and what are they? Are they disease-specific? What are the mitochondrial proteins involved in apoptosis in tissue X? Is leukemia linked to any neural disorder? What are the drugs targeting pathway Y?” and so forth can thus be addressed similarly.

**Conclusions.** We have integrated rich information associated with human TSGs from multiple sources in a web-accessible form to reveal many hidden links beyond tissue specificity. This makes VeryGene a potentially useful source for many applications, for instance, screening for therapeutic targets or biomarkers by tissue, subcellular localization, or gene-drug relationship or looking for functional enrichment of similarly localized genes or genes participating in a common pathway/disease or vice versa. And, most importantly, some hypotheses for pathogenic mechanism, diagnosis, and therapeutic research could be inferred based on the biological links of TSGs. Much of our effort will be geared toward the understanding of how TSGs play their roles in development, differentiation, stress response, and pathology. A study on tissue-specific transcriptional regulation is under way. We also expect to generate many testable hypotheses to maximize VeryGene’s practical value as a knowledge discovery tool.

Future development of VeryGene will aim to expand and update the extent of the current data set, ensure data quality control, and enhance user experience as well as data query capability to enable visualization of complex data relationships. In addition, expression profiles of diseased tissues will also be considered.

**ACKNOWLEDGMENTS.** We thank Dr. Changqing Zuo for insightful suggestions in design and implementation. We also appreciate Dr. Wei Liu’s continued support for this project. Our special thanks go to the authors and institutions of the named data sources used in building VeryGene.

**GRANTS.** This project is supported by a special grant from the Higher Education Agency of Guangdong Province, People’s Republic of China.

**DISCLOSURES.** No conflicts of interest, financial or otherwise, are declared by the author(s).

**REFERENCES.**


