This is a preprint of an article accepted for publication in *Physiological Genomics*. Yang X., Ye Y., Wang G., Huang H., Liang S. (2011). VeryGene: linking tissue-specific genes to diseases, drugs and beyond for knowledge discovery. Physiological Genomics , 43(8):457-460(doi:10.1152/physiolgenomics.00178.2010).

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], Shuang Liang[1]

[1]Institute of Genetic Engineering, Southern Medical University, Guangzhou, Guangdong Province, P.R. China

[2]Department of Biological and Chemical Engineering, Guangxi University of Technology, Liuzhou, Guangxi Province, P.R. China

[3]Guangzhou Medical College, Guangzhou, Guangdong Province, P.R. China

[4]School of Information, University of South Florida, Tampa, FL, USA

[5]Southern Medical University, Guangzhou, Guangdong Province, P.R. China

Corresponding author: Shuang Liang

Email: itshuang@gmail.com

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 such and additional information in order to speed up discovery and clinical practice. As an initiative toward systems biology, 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 and KEGG pathway, MGI Mammalian Phenotype, disease association and targeting drugs. The current release carefully selected 3960 annotated TSGs derived from 127 normal human tissues and cell types, including 5672 gene-disease and 2171 drug-target relationships. Other than 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.verygene.com.

Keywords: tissue specificity; disease; targeting drug

Human tissues exhibit distinct characteristics in spite of differentiating from a common origin to fulfil the different needs of our body. This kind of diversity is contributed largely by the coordinated expression of different tissue-specific genes, 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. Evidences gathered through mining tissue specificity, gene connectivity and disease association suggest that many disease-associated genes are likely to show specific expression in the tissues from which the diseases originate (9, 18). Furthermore, several studies had utilized tissue specificity as an important factor when characterizing therapeutic/drug targets (5, 27). Other areas for the use of tissue-specificity include, but not limited to, pathogenic mechanism, diagnosis, or therapeutic applications (17, 21, 23).

A number of databases have been created to facilitate studies of TSGs. For example, BioGPS (24), TiGER (11), COXPRESdb (15) and TiSGeD (26) databases can be used to query human gene expression in various tissues. However, most of the above 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 altogether, the above databases could not serve to the users' best interest alone. 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 the above information is highly desirable.

Herein, we present a web-accessible tissue-specific gene 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 (~4000 samples combined) and tissue types (~130 tissue types combined) from normal human subjects through microarray expression profiling analysis. Therefore, only these two datasets were selected for integration because of their extended coverage.

Analytically, searching for tissue-specific genes amounts to comparing gene expression over many tissue types. To determine the tissue distribution for a given gene i across K tissue types, there exists P = K(K - 1)/2 pair-wise comparisons for K tissue types. In our previous analyses (10), a modified Tukey-Kramer's honest significant difference (HSD) test with an Enrichment Score (equation 1) was proposed to overcome the type I error from multiple tests. A HSD test generates one Q value (difference of means between tissue pairs over standard deviation) per pair-wise comparison for each gene. An ES ($ES \in (0,1)$) takes into account of 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 type I error from multiple HSD tests, each and every TSGi was identified by observing an ES greater than that by chance alone (estimated by permutation). z_{ij}-scores (equation 2) were calculated to represent the relative level of a given TSGi expressed in one particular tissue j (j=1 to K) with regard to the mean expression of TSGi across all K tissues. The product of z_{ij} and ES_i, denoted as τ_{ij} (equation 3), was computed to account for both tissue specificity/selectivity and relative expression level of a TSG_i in a given tissue j. A large τ_{ij} specifies that a TSG_i 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).

$$ES_j = \frac{1}{P-1}\sum_{p=1}^{P}\left(1-\frac{Q_p - Min(Q_p)}{Max(Q_p) - Min(Q_p)}\right) \quad (1)$$

$$z_{ij} = \frac{x_{ij} - \mu_i}{\sigma_i} \quad (2)$$

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

Probe IDs were mapped to Entrez Gene IDs. Tissue names were carefully unified according to standard anatomical terms, and redundant tissue affiliations were merged according to the mean value of τ. Finally, 3960 tissue-specific genes were identified through expression profiling of a panel of 127 human tissue and cell types. These TSGs express selectively in ~ 2 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 annotation, biological pathways, mammalian phenotype linkage, disease association and targeting drugs. Subcellular localization information for these tissue-specific genes was retrieved from LOCATE (19), supplemented with cellular component annotation of the Gene Ontology (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 derived from MGI (Mouse Genome Informatics) (3). Gene-disease relationships were gathered from Gene2MeSH (1), OMIM information (7) and Swiss-Prot (14). Non-standard disease names were associated with MeSH IDs and mapped to the MeSH tree categories. Gene-targeting drug relationships were obtained from DrugBank (22). By integrating these data, 5672 gene-disease relationship and 2171 gene-drug relationship have been collected (Table 1).

Table 1.VeryGene data status.

Data Type	Number
Tissue/Cell types	127
TSGs	3960
TSG - Disease relationships	5672
TSG - Drug relationships	2171
TSG - Subcellular Localization	3687
TSG - GO annotation	47418
TSG - Pathway	6359
TSG - Mammalian Phenotype	32397

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, Gene Ontology, 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 is 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 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 dataset used to build VeryGene are downloadable for offline 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, 8 TSGs for periodontitis (MeSH:D010518) could be found from Multiple View. Batch View analysis shows that 5 genes among them are also

related to rheumatoid arthritis (MeSH:D001172). These 5 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 signalling 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). 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 (Figure 1) (16). Besides, it is well known that drug development is time-consuming and very expensive. Finding new indications of 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 eight most significant MeSH/MIM terms ($p = 0$). Most of these terms can be broadly classified into vascular or inflammatory diseases, among which Endometriosis (MeSH:D004715) (Figure 2) distinguishes itself from others as being suggested to be an autoimmune disease. The potential protective effect of Simvastatin on Endometriosis was preliminarily verified by study in nude mouse model (2). The above 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 it 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 up for functional enrichment of similarly localized genes or genes participates in a common pathway/disease or vice versa. And most importantly, some hypotheses for pathogenic mechanism, diagnosis and therapeutic researches, 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. 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 current dataset, ensure data quality control, and to enhance user experience as well as data query capability to enable visualization of the complex data relationship. Besides, expression profiles of diseased tissues will also be considered.

GRANT

This project is supported by a special grant from the Higher Education Agency of Guangdong

Province, P.R. China.

ACKNOWLEDGMENTS

We thank Dr. Changqing Zuo for insightful suggestions in the design and implementation. Our special thanks go to the authors and institutions of the named data sources used in building VeryGene. We are also grateful to the anonymous reviewers' suggestions for enhancing VeryGene's capabilities.

REFERENCES

1. **Ade A, Wright Z, and States D**. Gene2MeSH. *Ann Arbor (MI): National Center for IntergrativeBiomedical Informatics*. 2007 [The database is available at http://gene2mesh.ncibi.org/].

2. **Bruner-Tran KL, Osteen KG, and Duleba AJ**. Simvastatin Protects against the Development of Endometriosis in a Nude Mouse Model. *J Clin Endocrinol Metab* 94: 2489-2494, 2009.

3. **Bult CJ, Eppig JT, Kadin JA, Richardson JE, Blake JA, and the Mouse Genome Database Group**. The Mouse Genome Database (MGD): mouse biology and model systems. *Nucl Acids Res* 36: D724-728, 2008.

4. **Demmer RT, Behle JH, Wolf DL, Handfield M, Kebschull M, Celenti R, Pavlidis P, and Papapanou PN**. Transcriptomes in Healthy and Diseased Gingival Tissues. *Journal of Periodontology* 79: 2112-2124, 2008.

5. **Dezso Z, Nikolsky Y, Sviridov E, Shi W, Serebriyskaya T, Dosymbekov D, Bugrim A, Rakhmatulin E, Brennan R, Guryanov A, Li K, Blake J, Samaha R, and Nikolskaya T**. A comprehensive functional analysis of tissue specificity of human gene expression. *BMC Biology* 6: 49, 2008.

6. **Gene Ontology Consortium**. The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res* 32: D258-261, 2004.

7. **Hamosh A, Scott AF, Amberger JS, Bocchini CA, and McKusick VA**. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res* 33: D514-517, 2005.

8. **Kanehisa M, Goto S, Furumichi M, Tanabe M, and Hirakawa M**. KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic Acids Res* 38: D355-360, 2010.

9. **Lage K, Hansen NT, Karlberg EO, Eklund AC, Roque FS, Donahoe PK, Szallasi Z, Jensen TS, and Brunak S**. A large-scale analysis of tissue-specific pathology and gene expression of human disease genes and complexes. *Proceedings of the National Academy of Sciences of the United States of America* 105: 20870-20875, 2008.

10. **Liang S, Li Y, Be X, Howes S, and Liu W**. Detecting and profiling tissue-selective genes. *Physiol Genomics* 26: 158-162, 2006.

11. **Liu X, Yu X, Zack D, Zhu H, and Qian J**. TiGER: A database for tissue-specific gene expression and regulation. *BMC Bioinformatics* 9: 271, 2008.

12. **Matthews L, Gopinath G, Gillespie M, Caudy M, Croft D, de Bono B, Garapati P, Hemish J, Hermjakob H, Jassal B, Kanapin A, Lewis S, Mahajan S, May B, Schmidt E, Vastrik I, Wu G, Birney E, Stein L, and D'Eustachio P**. Reactome knowledgebase of human biological pathways and processes. *Nucleic Acids Res* 37: D619-622, 2009.

13. **Modi DK, Chopra VS, and Bhau U**. Rheumatoid arthritis and periodontitis: biological links and

the emergence of dual purpose therapies. *Indian J Dent Res* 20: 86-90, 2009.

14. **Mottaz A, Yip Y, Ruch P, and Veuthey A-L**. Mapping proteins to disease terminologies: from UniProt to MeSH. *BMC Bioinformatics* 9: S3, 2008.

15. **Obayashi T, Hayashi S, Shibaoka M, Saeki M, Ohta H, and Kinoshita K**. COXPRESdb: a database of coexpressed gene networks in mammals. *Nucleic Acids Res* 36: D77-82, 2008.

16. **Ortiz P, Bissada NF, Palomo L, Han YW, Al-Zahrani MS, Panneerselvam A, and Askari A**. Periodontal Therapy Reduces the Severity of Active Rheumatoid Arthritis in Patients Treated With or Without Tumor Necrosis Factor Inhibitors. *Journal of Periodontology* 80: 535-540, 2009.

17. **Pacak C, Sakai Y, Thattaliyath B, Mah C, and Byrne B**. Tissue specific promoters improve specificity of AAV9 mediated transgene expression following intra-vascular gene delivery in neonatal mice. *Genetic Vaccines and Therapy* 6: 13, 2008.

18. **Reverter A, Ingham A, and Dalrymple B**. Mining tissue specificity, gene connectivity and disease association to reveal a set of genes that modify the action of disease causing genes. *BioData Mining* 1: 8, 2008.

19. **Sprenger J, Lynn Fink J, Karunaratne S, Hanson K, Hamilton NA, and Teasdale RD**. LOCATE: a mammalian protein subcellular localization database. *Nucleic Acids Res* 36: D230-233, 2008.

20. **Su AI, Wiltshire T, Batalov S, Lapp H, Ching KA, Block D, Zhang J, Soden R, Hayakawa M, Kreiman G, Cooke MP, Walker JR, and Hogenesch JB**. A gene atlas of the mouse and human protein-encoding transcriptomes. *Proceedings of the National Academy of Sciences of the United States of America* 101: 6062-6067, 2004.

21. **Vasmatzis G, Klee EW, Kube DM, Therneau TM, and Kosari F**. Quantitating tissue specificity of human genes to facilitate biomarker discovery. *Bioinformatics* 23: 1348-1355, 2007.

22. **Wishart DS, Knox C, Guo AC, Cheng D, Shrivastava S, Tzur D, Gautam B, and Hassanali M**. DrugBank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Res* 36: D901-906, 2008.

23. **Wu C, Lin J, Hong M, Choudhury Y, Balani P, Leung D, Dang LH, Zhao Y, Zeng J, and Wang S**. Combinatorial Control of Suicide Gene Expression by Tissue-specific Promoter and microRNA Regulation for Cancer Therapy. *Molecular Therapy* 17: 2058-2066, 2009.

24. **Wu C, Orozco C, Boyer J, Leglise M, Goodale J, Batalov S, Hodge C, Haase J, Janes J, Huss J, and Su A**. BioGPS: an extensible and customizable portal for querying and organizing gene annotation resources. *Genome Biology* 10: R130, 2009.

25. **Wu G, Zhu L, Dent JE, and Nardini C**. A Comprehensive Molecular Interaction Map for Rheumatoid Arthritis. *PLoS ONE* 5: e10137, 2010.

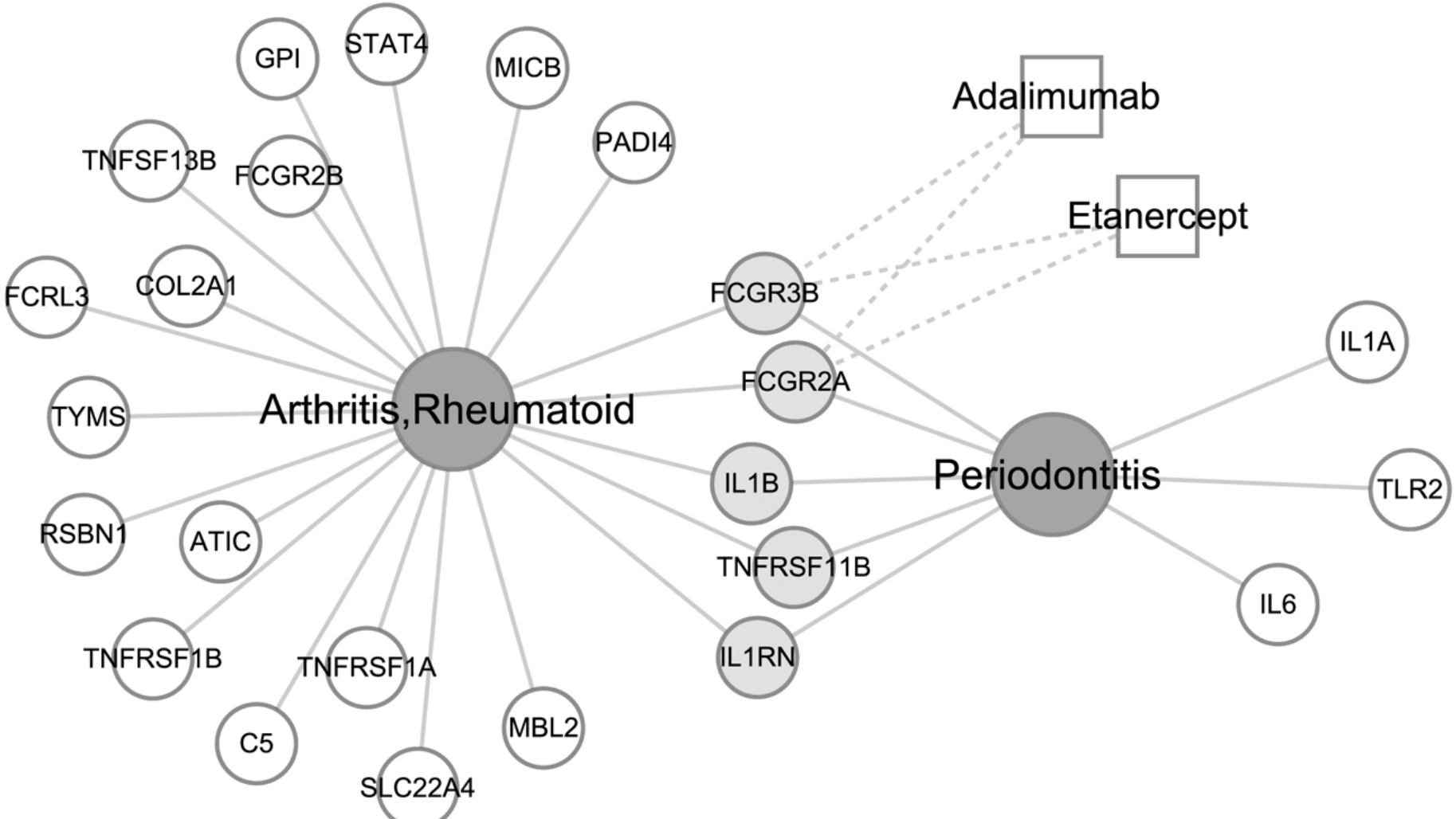
26. **Xiao S-J, Zhang C, Zou Q, and Ji Z-L**. TiSGeD: a database for tissue-specific genes. *Bioinformatics* 26: 1273-1275, 2010.

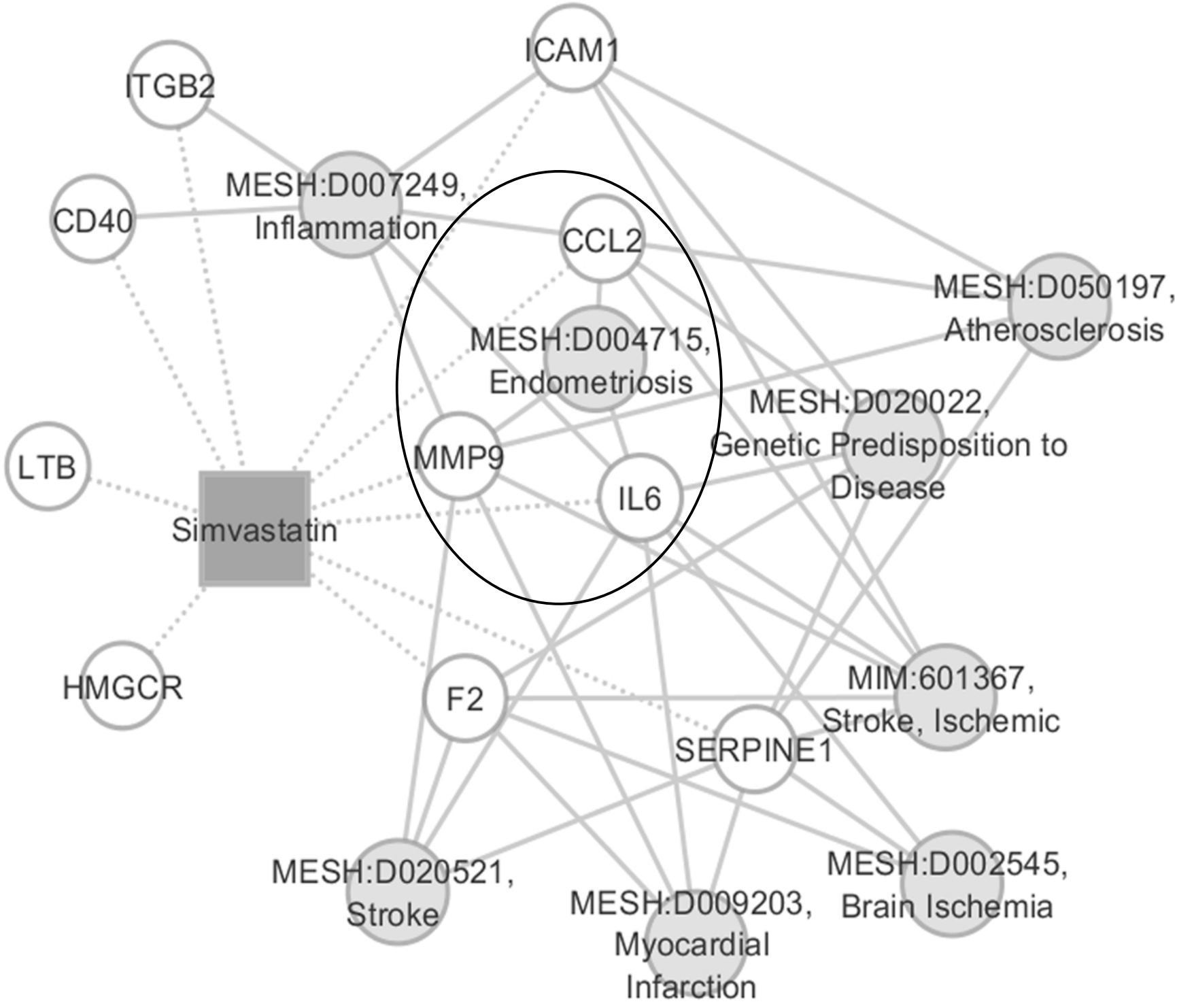
27. **Zheng CJ, Han LY, Yap CW, Ji ZL, Cao ZW, and Chen YZ**. Therapeutic targets: Progress of their exploration and investigation of their characteristics. *Pharmacological Reviews* 58: 259-279, 2006.

Fig. 1. An example of discovering TSG-targeting drugs by linking two different diseases, susceptible TSGs and common targeting drugs.

Illustration of a disease-TSG-drug network with Cytoscape. Different nodes are specified according to their categories as follows: disease, dark circle; drug, white square; common TSG, gray circle; other TSG, white circle. Lines are drawn according to their categories as follows: disease-TSG relationship, solid; drug-TSG relationship, dashed.

Fig. 2. An example of uncovering genes and diseases targeted by Simvastatin.
A filtered TSG-disease network visualized with Cytoscape where significant relationships ($p = 0$) are displayed. Nodes are colored and shaped according to their categories as follows: disease, gray circle; drug, dark square; TSG, white circle. Lines are drawn according to their categories as follows: disease-TSG relationship, solid; drug-TSG relationship, dashed. Simvastatin-targeting TSGs in the subset are related to Endometriosis and several other diseases.


GPI
STAT4
MICB
PADI4
TNFSF13B
FCGR2B
FCRL3
COL2A1
TYMS
Arthritis,Rheumatoid
RSBN1
ATIC
TNFRSF1B
TNFRSF1A
C5
MBL2
SLC22A4
Adalimumab
Etanercept
FCGR3B
FCGR2A
IL1B
TNFRSF11B
IL1RN
Periodontitis
IL1A
TLR2
IL6


ICAM1
ITGB2
MESH:D007249,
Inflammation
CD40
CCL2
MESH:D050197,
Atherosclerosis
MESH:D004715,
Endometriosis
MESH:D020022,
Genetic Predisposition to
Disease
MMP9
LTB
IL6
Simvastatin
HMGCR
F2
MIM:601367,
Stroke, Ischemic
SERPINE1
MESH:D020521,
Stroke
MESH:D009203,
Myocardial
Infarction
MESH:D002545,
Brain Ischemia