Text-mining applied to autoimmune disease research: the Sjögren’s syndrome knowledge base

Sven-Ulrik Gorr1*, Trevor J Wennblom2, Steve Horvath3, David TW Wong4 and Sara A Michie5

Abstract

Background: Sjögren’s syndrome is a tissue-specific autoimmune disease that affects exocrine tissues, especially salivary glands and lacrimal glands. Despite a large body of evidence gathered over the past 60 years, significant gaps still exist in our understanding of Sjögren’s syndrome. The goal of this study was to develop a database that collects and organizes gene and protein expression data from the existing literature for comparative analysis with future gene expression and proteomic studies of Sjögren’s syndrome.

Description: To catalog the existing knowledge in the field, we used text mining to generate the Sjögren’s Syndrome Knowledge Base (SSKB) of published gene/protein data, which were extracted from PubMed using text mining of over 7,700 abstracts and listing approximately 500 potential genes/proteins. The raw data were manually evaluated to remove duplicates and false-positives and assign gene names. The data base was manually curated to 477 entries, including 377 potential functional genes, which were used for enrichment and pathway analysis using gene ontology and KEGG pathway analysis.

Conclusions: The Sjögren’s syndrome knowledge base (http://sskb.umn.edu) can form the foundation for an informed search of existing knowledge in the field as new potential therapeutic targets are identified by conventional or high throughput experimental techniques.

Background

Sjögren’s syndrome is a tissue-specific autoimmune disease that affects exocrine tissues, especially salivary glands and lacrimal glands. It is one of the most common autoimmune disorders in the U.S., with an estimated prevalence of 2–4 million people. The autoimmunity-mediated damage of the salivary and lacrimal glands in Sjögren’s syndrome leads to a decrease in the production of saliva and tears and to the development of dry mouth and dry eyes. Without the lubricating and protective functions of saliva and tears, the oral and ocular surfaces are subject to infections and discomfort, leading to a significantly reduced quality of life [1,2].

Development of Sjögren’s syndrome requires a complex interplay between a number of genetic, hormonal and environmental factors, most of which have not been defined. Genetic linkages, especially involving major histocompatibility complex (MHC) genes, have been reported for Sjögren’s syndrome but it is not clear if, or how, the associated genes are involved in the development of the disease [3]. Additional non-MHC genes have also been linked with the development of Sjögren’s syndrome.

In addition to genetic predisposition, some studies suggest that infection of a genetically-susceptible individual by a virus or other pathogen might trigger the development of an autoimmune disease [4]. The proposed mechanisms include activation of the innate immune system, release of self antigens from damaged or apoptotic tissues, and molecular mimicry that results in activation of T cells and/or B cells that react with tissue antigens [4].

Both the innate and the adaptive immune systems are involved in the pathogenesis of Sjögren’s syndrome. The type I interferon (IFN) pathway, which plays an important role in the innate immune response to viruses, is also thought to play an important role in the development of Sjögren’s syndrome and other autoimmune disorders, including SLE [5,6]. Moreover, type I IFNs can...
activate the adaptive immune system directly, by binding to IFN receptors on antigen presenting cells, T cells and B cells, or indirectly, by inducing the production and release of cytokines and chemokines that bind to these cells.

Autoantibodies to intracellular antigens, notably the nuclear proteins SSA/Ro and SSB/La, are found in the sera of many patients with Sjögren’s syndrome. These autoantibodies are thought to develop when intracellular antigens, some of which have undergone proteolytic cleavage that reveals new antigenic epitopes, become “visible” to the immune system in membrane blebs on the surface of apoptotic cells [7]. Alternatively, antigenic epitopes from bacteria and viruses, including Epstein-Barr virus (EBV) and coxsackie virus, may act as molecular mimics that trigger the development of antibodies that cross react with similar epitopes on target tissue autoantigens [8,9]. Although autoantibodies to intracellular antigens are useful in the diagnosis of Sjögren’s syndrome, it is not clear if they play a direct role in the development of salivary gland and lacrimal gland damage and hypofunction. In contrast, autoantibodies to the M3 muscarinic acetylcholine receptor (M3R) have been directly implicated in salivary gland hypofunction in the nonobese diabetic (NOD) mouse model of Sjögren’s syndrome [10]. Importantly, function-inhibiting anti-M3R autoantibodies are found in the sera of many patients with Sjögren’s syndrome [11].

Current therapy for Sjögren’s syndrome usually consists of palliative treatment that relieves the symptoms of dry eye and dry mouth, but fails to modify the underlying disease. Novel disease-modifying treatment strategies, based on recent immunological insights in Sjögren’s syndrome and other autoimmune diseases, have met with mixed results [12]. For example, in recent clinical trials, treatment of Sjögren’s syndrome patients with a B cell-depleting anti-CD20 monoclonal antibody (rituximab) led to significant improvement of the stimulated whole saliva flow rate and a reduction in parotid and proteomic studies have identified many genes and pathways that may play a role in the pathogenesis of Sjögren’s syndrome [15-17]. However, validation of these data will require significant additional effort. As an initial step in this validation, we have compiled the published data on Sjögren’s syndrome that is not derived from gene expression or proteomic studies. No such unifying database currently exists. Through data curation, the existing data have been uniformly formatted to allow systematic retrieval and comparisons to newly generated gene expression data. As an example of its functionality, the Sjögren’s Syndrome Knowledge Base (SSKB) was analyzed for biological functions and pathways that are likely to play a role in the disease.

Construction and content

Data mining
To catalog the existing knowledge in the field, we used text mining to generate the Sjögren’s Syndrome Knowledge Base (SSKB) of published gene/protein data (http://sskb.umn.edu/) [18]. The focus of this data-base is on individually identified genes and proteins. Thus, microarray experiments were not included. The raw data for SSKB was extracted from PubMed [19]) using the text mining program EBIMed (http://www.ebi.ac.uk/Rehholz-srv/ebimed/) [20] with the search term “Sjögren’s Syndrome” restricted to “MeshHeadingsList”. The foundational search identified over 7,700 abstracts and approximately 500 potential genes/proteins. The SSKB is continually updated by regular automated searches of PubMed followed by manual curation.

Curation of raw data
The identified abstracts were manually evaluated to remove duplicates and false-positives. In older publications, where gene names were not readily identifiable, names were assigned based on in depth evaluation of the protein name context and available gene data in public databases, including the National Center for Biotechnology Information’s Entrez search engine [21] and UniProt [22,23]. The SSKB includes data from human studies and animal models. For the genes identified in animal models, the human homolog was identified by automated ortholog search, using WebGestalt 2.0 [24,25]. These steps reduced the database to 477 current entries. The online database contains the fully curated data and currently contains 413 entries, which can be accessed at http://sskb.umn.edu. Updates and newly curated data are continually added.

The 477 entries were sorted to identify autoantigens and viral/bacterial antigens, resulting in 377 potential functional genes, which were used for enrichment and pathway analysis.

Enrichment analysis
The 377 human gene entries were used for subsequent enrichment analyses in Webgestalt [24,25]. Gene enrichment in the SSKB gene set was compared to the human
genome using the hypergeometric test with multiple test adjustment [26] and a significance level of P < 0.01.

The Gene Ontology [27,28] was accessed with Webgestalt and analysis was restricted to processes and functions represented by two or more genes. Pathway analysis was performed with Webgestalt in the Kyoto Encyclopedia of Genes and Genomes (KEGG) [29,30]. The selection was restricted to pathways with 4 or more genes represented, resulting in identification of 72 KEGG pathways. The "salivary secretion" pathway (KO04970) was recently added to KEGG (11/9/10) and was not included in this analysis. This pathway contains 59 genes, seven of which are found in the SSKB gene set.

Utility and discussion
We constructed a database containing proteins and genes associated with Sjögren’s syndrome in human disease or animal models, as identified by text mining of published data. The public SSKB currently contains 413 genes/proteins and can be viewed online (http://sskb.umn.edu/). All genes have been assigned gene symbols and UniProt IDs, which allows rapid retrieval of gene-specific data from external databases. The SSKB database can be used to determine whether a list of genes is enriched with known Sjögren’s syndrome genes and one can carry out a function enrichment analysis (hypergeometric distribution). Individual genes and the corresponding gene products, synonyms and alternate names can be searched by using a web browser search function. Autoantigens, viral antigens and bacterial antigens are separately identified under “Antigens”. The SSKB is continually maintained and updated and new genes are added as their analysis is completed.

Based on the abstracts used to retrieve the SSKB genes/proteins, 85 proteins were initially characterized as autoantigens and 15 proteins were characterized as viral (14) or bacterial (1) antigens. Not surprisingly, SSA/Ro and SSB/La were among the most frequently retrieved autoantigens. It has been proposed that viral or bacterial antigens act as autoimmune triggers by molecular mimicry of endogenous human proteins [2,8,9]. However, eight of the 14 putative viral antigens in SSKB were selected for BLAST analysis, which did not identify strong sequence similarity with human proteins (not shown).

The 377 proteins not identified as autoantigens or microbial antigens were considered candidates for functional genes that could play a role in the initiation and progression of Sjögren’s syndrome. Since the gene list contains data from humans and animals, the corresponding human genes were identified, with the assumption that genes identified in animal models of Sjögren’s syndrome may also be involved in the human disease.

Gene ontology
The Gene Ontology database [27] was queried to identify the biological processes, cellular components and molecular functions associated with genes in the SSKB (Table 1). The 40 most highly enriched entries were identified in each category.

The most highly enriched biological processes (19 of 40; 18 of the top 20) were associated with immune function, including leukocyte proliferation, leukocyte activation, and regulation of the immune response. Other prominent biological processes were associated with apoptosis and cell death. Thus, the SSKB data set is consistent with recent microarray data [16] and reflects current models for the biological processes involved in the pathogenesis of Sjögren’s syndrome [5,31,32].

The most highly enriched cellular component was the calcineurin complex, which plays a major role in the activation of T cells. Interestingly, in placebo-controlled clinical trials, treatment of Sjögren’s syndrome patients with eye drops that contain the calcineurin inhibitor cyclosporine, led to significant improvement in several of the signs and symptoms of dry eye [33].

Other highly enriched cellular components include: 1) platelet alpha granules. Although platelet activation has been reported in the salivary glands of Sjögren’s syndrome patients [34], a direct search of PubMed for “platelet alpha granules” with “sjogren’s” did not retrieve any published studies. Thus, while the proteins identified were retrieved from the literature, their potential association with platelet alpha granules in Sjögren’s syndrome has not previously been noted. 2) MHC protein complexes were identified and are presumably involved in the presentation of autoantigens [16]. 3) The finding that protein-lipid complexes and lipoprotein particles are associated with Sjögren’s syndrome may be consistent with changes in serum lipid levels in Sjögren’s syndrome patients [35] although the prevalence of anti-phospholipid antibodies is low in Sjögren’s syndrome [36]. 4) Nerve terminals and axons were also prominent cellular components, consistent with the known neurological component of Sjögren’s syndrome [37].

In molecular function, nitric oxide synthase (NOS) activity was the most highly enriched, although only three genes (NOS1-3) were identified. Nitric oxide (NO) signaling appears to be directly affected in salivary and lacrimal glands in Sjögren’s syndrome [38]. Other highly enriched molecular functions include chemokine and cytokine activity/receptor binding (8 of the top 15) and peptidase activities.

Pathway analysis
The SSKB gene list was submitted to KEGG [29] to identify biological pathways potentially associated with Sjögren’s syndrome. A total of 72 KEGG pathways
### Table 1 Gene Ontology enrichment analysis

| Rank | BIOLOGICAL PROCESS                                                                 | GO ID    | Reference Genes | Observed Genes | Ratio     |
|------|-------------------------------------------------------------------------------------|----------|-----------------|----------------|-----------|
| 1    | regulation of lymphocyte proliferation                                             | GO:0050670 | 81              | 32             | 39.51%    |
| 2    | regulation of leukocyte proliferation                                               | GO:0070663 | 82              | 32             | 39.02%    |
| 3    | regulation of mononuclear cell proliferation                                        | GO:0032944 | 82              | 32             | 39.02%    |
| 4    | adaptive immune response based on somatic recombination of immune receptors built from immunoglobulin superfamily domains | GO:0002460 | 112             | 38             | 33.93%    |
| 5    | lymphocyte proliferation                                                            | GO:0046651 | 112             | 37             | 33.04%    |
| 6    | leukocyte proliferation                                                             | GO:0070661 | 114             | 37             | 32.46%    |
| 7    | mononuclear cell proliferation                                                      | GO:0032943 | 114             | 37             | 32.46%    |
| 8    | regulation of lymphocyte activation                                                | GO:0051249 | 141             | 42             | 29.79%    |
| 9    | regulation of cell activation                                                      | GO:0050865 | 168             | 46             | 27.38%    |
| 10   | regulation of leukocyte activation                                                 | GO:0002694 | 159             | 43             | 27.04%    |
| 11   | positive regulation of immune system process                                        | GO:0002684 | 229             | 60             | 26.20%    |
| 12   | regulation of immune response                                                      | GO:0050776 | 218             | 54             | 24.77%    |
| 13   | lymphocyte activation                                                              | GO:0046649 | 272             | 59             | 21.69%    |
| 14   | leukocyte activation                                                               | GO:0045321 | 324             | 66             | 20.37%    |
| 15   | inflammatory response                                                              | GO:0006954 | 359             | 71             | 19.78%    |
| 16   | cell activation                                                                    | GO:0001775 | 366             | 71             | 19.40%    |
| 17   | immune response                                                                    | GO:0006955 | 750             | 133            | 17.73%    |
| 18   | regulation of response to stimulus                                                 | GO:0048583 | 441             | 75             | 17.01%    |
| 19   | defense response                                                                   | GO:0006952 | 657             | 100            | 15.22%    |
| 20   | immune system process                                                              | GO:0002376 | 1066            | 162            | 15.20%    |
| 21   | response to wounding                                                               | GO:0009611 | 560             | 85             | 15.18%    |
| 22   | response to external stimulus                                                      | GO:0009605 | 904             | 110            | 12.17%    |
| 23   | multi-organism process                                                             | GO:0051704 | 668             | 79             | 11.83%    |
| 24   | regulation of programmed cell death                                                | GO:0043067 | 812             | 92             | 11.33%    |
| 25   | regulation of apoptosis                                                             | GO:0042981 | 805             | 91             | 11.30%    |
| 26   | regulation of cell death                                                            | GO:0010941 | 815             | 92             | 11.29%    |
| 27   | regulation of cell proliferation                                                   | GO:0042127 | 739             | 79             | 10.69%    |
| 28   | apoptosis                                                                          | GO:0006915 | 1063            | 102            | 9.60%     |
| 29   | programmed cell death                                                              | GO:0012501 | 1071            | 102            | 9.52%     |
| 30   | response to chemical stimulus                                                      | GO:0042221 | 1243            | 117            | 9.41%     |
| 31   | cell proliferation                                                                  | GO:0008283 | 1056            | 98             | 9.28%     |
| 32   | death                                                                               | GO:0016265 | 1171            | 107            | 9.14%     |
| 33   | cell death                                                                          | GO:0008219 | 1167            | 106            | 9.08%     |
| 34   | response to stress                                                                 | GO:0006950 | 1696            | 144            | 8.49%     |
| 35   | positive regulation of biological process                                          | GO:0048518 | 1865            | 153            | 8.20%     |
| 36   | positive regulation of cellular process                                            | GO:0048522 | 1699            | 130            | 7.65%     |

| Rank | CELLULAR COMPONENT            | GO ID    | Count | Observed | Ratio  |
|------|-------------------------------|----------|-------|----------|--------|
| 1    | calcineurin complex           | GO:0005955 | 5 | 3 | 60.00% |
| 2    | external side of plasma membrane | GO:0009897 | 131 | 40 | 30.53% |
Table 1 Gene Ontology enrichment analysis (Continued)

| Rank | MOLECULAR FUNCTION | GO ID   | COUNT | Observed | RATIO  |
|------|--------------------|---------|-------|----------|--------|
| 1    | arginine binding   | GO:0034618 | 3    | 3        | 100.00%|
| 2    | nitric-oxide synthase activity | GO:0004517 | 3    | 3        | 100.00%|
| 3    | tetrahydrobiopterin binding | GO:0034617 | 3    | 3        | 100.00%|
| 4    | C-X-C chemokine binding | GO:0019958 | 8    | 4        | 50.00% |
| 5    | beta-amyloid binding | GO:0001540 | 13   | 5        | 38.46% |
| 6    | tumor necrosis factor receptor binding | GO:0005164 | 21   | 8        | 38.10% |
showed highly significant enrichment (P < 0.001) in this analysis (Table 2).

The pathway analysis revealed dominant pathways associated with immune regulation. Indeed, the eight most highly enriched pathways were associated with antigen presenting cells and activation of T cells and B cells.

Several cancer associated pathways were identified. This is partly due to the overlap between cancer pathways. These pathways typically include cytokine or growth factor stimulation of cell cycle and cell death and were not further analyzed.

Pathways associated with apoptosis, cytokine signaling and inflammation were also highly enriched. To focus on the events associated with initiation of Sjögren's syndrome, we analyzed pathways with known triggers. Several of the highly enriched pathways are triggered by bacterial toxins, viral DNA, or viral RNA. These include signaling pathways for Toll-like receptor, NOD-like receptor, RIG-I-

Table 1 Gene Ontology enrichment analysis (Continued)

| Gene Ontology ID | Observed Genes | Reference Genes | Ratio |
|------------------|----------------|-----------------|-------|
| GO:0008009       | 47             | 17              | 36.17%|
| GO:0042379       | 49             | 17              | 34.69%|
| GO:0015026       | 19             | 6               | 31.58%|
| GO:0032813       | 31             | 9               | 29.03%|
| GO:0005126       | 178            | 46              | 25.84%|
| GO:0019956       | 26             | 6               | 23.08%|
| GO:0005125       | 196            | 45              | 22.96%|
| GO:0070851       | 67             | 14              | 20.90%|
| GO:0005518       | 35             | 7               | 20.00%|
| GO:001664        | 107            | 20              | 18.69%|
| GO:0005178       | 58             | 9               | 15.52%|
| GO:0004197       | 71             | 10              | 14.08%|
| GO:0008083       | 161            | 19              | 11.80%|
| GO:0019955       | 108            | 12              | 11.11%|
| GO:0046982       | 189            | 21              | 11.11%|
| GO:0005539       | 139            | 14              | 10.07%|
| GO:0032403       | 196            | 19              | 9.69% |
| GO:0005102       | 856            | 83              | 9.70% |
| GO:0005057       | 159            | 15              | 9.43% |
| GO:0001871       | 153            | 14              | 9.15% |
| GO:0030414       | 154            | 14              | 9.09% |
| GO:0030246       | 349            | 29              | 8.31% |
| GO:0004175       | 370            | 28              | 7.57% |
| GO:0030247       | 153            | 14              | 9.15% |
| GO:0046983       | 514            | 36              | 7.00% |
| GO:0042802       | 618            | 38              | 6.15% |
| GO:0019899       | 505            | 29              | 5.74% |
| GO:0008233       | 563            | 30              | 5.33% |
| GO:0070011       | 546            | 29              | 5.31% |
| GO:0060089       | 2116           | 98              | 4.63% |
| GO:0004871       | 2116           | 98              | 4.63% |
| GO:0004872       | 1674           | 71              | 4.24% |
| GO:0005515       | 8041           | 280             | 3.48% |
| GO:0005488       | 12465          | 320             | 2.57% |

The table ranks the gene enrichment in biological processes, cellular component and molecular function with corresponding GO IDs. For each GO ID, the number of Observed Genes identified in the SSKB was divided by the number of Reference Genes in the human genome to calculate the Ratio of enrichment (Ratio).
### Table 2 Biological pathways associated with SSKB genes

| Rank | PATHWAY                                                   | SSKB Genes | ENRICHMENT | Raw P       | Adjust P     |
|------|----------------------------------------------------------|------------|------------|-------------|--------------|
| 1    | Allograft rejection                                      | 23         | 76.02      | 3.62E-39    | 6.82E-38     |
| 2    | Intestinal immune network for IgA production             | 27         | 67.82      | 7.26E-44    | 2.05E-42     |
| 3    | Asthma                                                   | 14         | 58.61      | 4.14E-22    | 2.75E-21     |
| 4    | Type I diabetes mellitus                                | 20         | 57.09      | 9.13E-31    | 9.38E-30     |
| 5    | Graft-versus-host disease                               | 18         | 53.83      | 3.21E-27    | 2.79E-26     |
| 6    | Autoimmune thyroid disease                              | 22         | 52.13      | 1.29E-32    | 1.82E-31     |
| 7    | Primary immunodeficiency                                | 14         | 50.24      | 6.38E-21    | 3.79E-20     |
| 8    | Hematopoietic cell lineage                              | 33         | 47.1       | 1.39E-46    | 5.24E-45     |
| 9    | Toll-like receptor signaling pathway                    | 37         | 46.01      | 1.13E-51    | 6.38E-50     |
| 10   | Apoptosis                                                | 25         | 35.68      | 5.55E-32    | 6.97E-31     |
| 11   | NOD-like receptor signaling pathway                     | 17         | 34.44      | 7.61E-22    | 4.78E-21     |
| 12   | Amyotrophic lateral sclerosis (ALS)                     | 14         | 33.18      | 5.81E-18    | 2.85E-17     |
| 13   | Other glycan degradation                                | 4          | 31.4       | 6.67E-06    | 1.24E-05     |
| 14   | Cytokine-cytokine receptor interaction                  | 66         | 31.05      | 5.91E-79    | 6.68E-77     |
| 15   | T cell receptor signaling pathway                       | 26         | 30.24      | 4.12E-31    | 4.66E-30     |
| 16   | RIG-I-like receptor signaling pathway                   | 17         | 30.07      | 9.98E-21    | 5.64E-20     |
| 17   | Cell adhesion molecules (CAMs)                          | 32         | 29.99      | 6.40E-38    | 1.03E-36     |
| 18   | Bladder cancer                                          | 10         | 29.9       | 1.06E-12    | 3.24E-12     |
| 19   | Viral myocarditis                                       | 17         | 29.25      | 1.68E-20    | 9.04E-20     |
| 20   | Cytosolic DNA-sensing pathway                           | 13         | 29.16      | 5.78E-16    | 2.42E-15     |
| 21   | Pancreatic cancer                                       | 15         | 26.17      | 1.88E-17    | 8.50E-17     |
| 22   | Small cell lung cancer                                  | 16         | 23.92      | 7.32E-18    | 3.45E-17     |
| 23   | Glycosaminoglycan degradation                           | 4          | 23.92      | 2.13E-05    | 3.65E-05     |
| 24   | Natural killer cell mediated cytotoxicity               | 25         | 22.92      | 1.06E-26    | 8.56E-26     |
| 25   | ErbB signaling pathway                                  | 13         | 22.16      | 2.51E-13    | 8.86E-13     |
| 26   | Epithelial cell signaling in Helicobacter pylori infection | 12       | 22.16      | 2.64E-13    | 9.04E-13     |
| 27   | Complement and coagulation cascades                     | 12         | 21.84      | 3.17E-13    | 1.05E-12     |
| 28   | B cell receptor signaling pathway                       | 13         | 21.77      | 3.38E-14    | 1.23E-13     |
| 29   | Prion diseases                                          | 6          | 21.53      | 3.27E-07    | 6.84E-07     |
| 30   | Antigen processing and presentation                     | 15         | 21.17      | 5.49E-16    | 2.39E-15     |
| 31   | Colorectal cancer                                       | 14         | 20.93      | 6.14E-15    | 2.48E-14     |
| 32   | Adipocytokine signaling pathway                         | 11         | 20.62      | 6.05E-12    | 1.80E-11     |
| 33   | Chemokine signaling pathway                             | 30         | 19.83      | 7.80E-30    | 7.35E-29     |
| 34   | Prostate cancer                                         | 14         | 19.76      | 1.42E-14    | 5.33E-14     |
| 35   | Glioma                                                  | 10         | 19.32      | 1.10E-10    | 2.89E-10     |
| 36   | Jak-STAT signaling pathway                              | 23         | 18.64      | 1.67E-22    | 1.18E-21     |
| 37   | Non-small cell lung cancer                              | 8          | 18.61      | 1.13E-08    | 2.50E-08     |
| 38   | Melanoma                                                | 10         | 17.69      | 2.71E-10    | 6.96E-10     |
| 39   | Pathways in cancer                                      | 46         | 17.51      | 9.85E-43    | 2.23E-41     |
| 40   | Fc epsilon RI signaling pathway                         | 11         | 17.49      | 3.90E-11    | 1.05E-10     |
| 41   | Chronic myeloid leukemia                                | 10         | 16.75      | 4.74E-10    | 1.19E-09     |
| 42   | GnRH signaling pathway                                  | 12         | 14.92      | 3.42E-11    | 9.43E-11     |
| 43   | Leukocyte transendothelial migration                    | 14         | 14.9       | 7.91E-13    | 2.48E-12     |
| 44   | VEGF signaling pathway                                  | 9          | 14.87      | 1.04E-08    | 2.35E-08     |
like receptor signaling pathways and the cytosolic DNA-sensing pathway.

**Overlap with other autoimmune diseases**
The KEGG pathways include several pathways for autoimmune diseases, including type 1 diabetes mellitus, autoimmune thyroid disease, and SLE. While about 50% of the genes associated with the first two pathways are also associated with Sjögren's syndrome, only 16 Sjögren's syndrome genes were identified in the 140-gene SLE pathway (KEGG ID: hsa05322). These findings suggest that significant differences exist in the pathogenesis of autoimmune diseases.

**Conclusions**
The results of this analysis can serve as a background and comparison for the increasing number of gene expression data sets available for Sjögren's syndrome, e.g. [15-17]. Preliminary analysis of such data sets suggest that the biological pathways identified in the SSKB are very similar to those identified in human parotid tissue but quite different from those identified in human labial salivary glands [15]. Future analyses will further define these differences and focus on the comparison of biological pathways identified in human tissues and mouse models of Sjögren's syndrome. It is envisioned that the SSKB data can also serve as the starting point for literature reviews and literature-based validation of identified genes; functional gene enrichment studies; protein-protein interaction networks and other bioinformatics analyses; it can be used to arrive at gene sets for SNP set enrichment analysis (pathway based GWAS studies); it can be used to define a gene set for gene set enrichment analysis (GSEA); as a starting point for bioinformatics analysis.
protein-protein interaction networks (based on yeast 2 hybrid) can be identified among the SSKB genes.

**Availability and requirements**

The Sjögren’s syndrome knowledge base is freely available at sskb.umn.edu.

**Competing interests**

Dr. David Wong is scientific advisor to RNAmeTRIX Inc., a molecular diagnostics company. The authors declare no conflicts of interest.

**Authors’ contributions**

SUG collected and organized data and performed data analysis and drafted the manuscript. TW designed and implemented the database and web site. DTW contributed to data analysis, critical review of the database, statistical analysis and editing the manuscript. SM contributed to data analysis, critical review of the database and drafting the manuscript. All authors read and approved the final manuscript.

**Acknowledgements**

The authors thank Dr. Ammon Peck, University of Florida Dr. Michael Zhou, University School of Medicine, Stanford, CA 94305, USA. 3Department of Biostatistics, School of Public Health, University of Minnesota, Minneapolis, MN 55455, USA. 4School of Dentistry, University of California, Los Angeles, CA 90095, USA. 5Department of Pathology, Stanford University School of Medicine, Stanford, CA 94305, USA.

Received: 10 October 2011 Accepted: 18 June 2012

**Published** 3 July 2012

**References**

1. Meijer JM, Meiners PM, Huddleston Slater JJ, Spilkeret FK, Kallenberg CG, Vissink A, Bootema H: Health-related quality of life, employment and disability in patients with Sjögren’s syndrome. Rheumatology 2009, 48(9):1077–1082.
2. Rhodus NL: Sjögren’s syndrome. Quintessence Int 1999, 30(10):699–699.
3. Williams PH, Cobb BL, Namjoo B, Scafﬁeld RH, Sawalha AH, Harley JB: Horizons in Sjögren’s syndrome genetics. Clin Rev Allergy Immunol 2007, 32(3):201–209.
4. Ercolini AM, Miller SD: The role of infections in autoimmune disease. Clin Exp Immunol 2009, 155(1):1–15.
5. Mavragani CP, Crow MK: Activation of the type 1 interferon pathway in primary Sjogren’s syndrome. J Autoimmun 2010, 35(3):225–231.
6. Peck AB, Nguyen CQ, Sharma A, McIndoe RA, She JX: The Interferon-Signature of Sjögren’s Syndrome: What Does It Say about the etiopathology of autoimmunity? J Clin Rheumatol Musculoskel Med 2011, 3(1).
7. Rosen A, Cascio-Rosen L: Altered autoantigen structure in Sjögren’s syndrome: implications for the pathogenesis of autoimmune tissue damage. Crit Rev Oral Biol Med 2004, 15(3):156–164.
8. Stathopoulos EA, Rautjas JS, Shia EA, Moutsopoulos HM, Tsiaousis AG: Cross-reaction between antibodies to the major epitope of Ro60 KDa autoantigen and a homologous peptide of Cosackie virus 2B protein. Clin Exp Immunol 2005, 141(1):148–154.
9. Zhao Z, Granucci F, Yeh L, Schaffer PA, Cantor H: Molecular mimicry by herpes simplex virus type 1: autoimmune disease after viral infection. Science 1998, 279(5353):1344–1347.
10. Nguyen KH, Brayer J, Cha S, Diggas D, Yasunari U, Hill G, Peck AB, Humphreys-Behr MG: Evidence for antimuscarinic acetylcholine receptor antibody-mediated secretory dysfunction in NOD mice. Arthritis Rheum 2000, 43(10):2297–2306.
11. Li J, Ha Y, Ku N, Choi S, Lee S, Oh S, Kim J, Lee J, Lee E, Song Y, et al: Inhibitory effects of autoantibodies on the muscarinic receptors in Sjögren’s syndrome. Lab Invest 2004, 84(11):1430–1438.
12. von Bultzingglöwen I, Sollecito TP, Fox PC, Daniels T, Jonsson R, Lockhart PB, Way D, Brennan MT, Carozzo M, Gardner B, et al: Salivary dysfunction associated with systemic diseases: systematic review and clinical management recommendations. Oral Surg Oral Med Oral Pathol Oral Radiol Endod 2007, 103(Suppl):S75.e1–S75.e15.
13. Meijer JM, Meiners PM, Vissink A, Spilkeret FK, Abdulahad W, Kammenga N, Brouwer E, Kallenberg CG, Bootsmma H: Effectiveness of rituximab treatment in primary Sjögren’s syndrome: a randomized, double-blind, placebo-controlled trial. Arthritis Rheum 2010, 62(8):960–968.
14. Moutsopoulos NM, Katifis GE, Angelov N, Leakan RA, Sankar P, Pillemer S, Wahl SM: Lack of efﬁcacy of etanercept in Sjögren’s syndrome correlates with failed suppression of TNF(Alpha) and systemic immune activation. Ann Rheum Dis 2008, 67(10):1437–1443.
15. Gottenberg JE, Cagnard N, Lucchi N, Letourneur F, Mistou S, Lazure T, Jacques S, Ba N, Iltah M, Lucque C, et al: Activation of IFN pathways and plasmacytoid dendritic cell recruitment in target organs of primary Sjögren’s syndrome. Proc Natl Acad Sci U S A 2006, 103(8):2770–2775.
16. Hu S, Zhou M, Jiang J, Elashoff D, Gorr S, Michie SA, Spilkeret FK, Bootsmma H, Kallenberg CG, et al: Systems biology analysis of Sjögren’s syndrome and mucocutaneous lymphoid tissue lymphoma in parotid glands. Arthritis Rheum 2009, 60(1):81–92.
17. Nguyen CQ, Sharma A, Lee BH, She JX, McIndoe RA, Peck AB: Differential gene expression in the salivary gland during development and onset of xerostomia in Sjögren’s syndrome-like disease of the C57BL/6-NOD-Aec1Aec2 mouse. Arthritis Res Ther 2009, 11(2):R56.
18. Sjögren’s Syndrome Knowledge Base. http://sskb.umn.edu/.
19. PubMed. http://www.ncbi.nlm.nih.gov/pubmed/.
20. Reinhob-Schumann D, Kirch N, Arengu M, Gaudian S, Rietherow M, Stoehr P: EBiMed—text crunching to gather facts for proteins from Medline. Bioinformatics 2007, 23(23):237–244.
21. Entrez. http://www.ncbi.nlm.nih.gov/Entrez/.
22. UniProt. http://www.uniprot.org/.
23. Consortium TU: Ongoing and future developments at the Universal Protein Resource. Nuclic Acids Res 2011, 39(Suppl 1):D214–D219.
24. WebGestalt. http://bioinfo.vanderbilt.edu/webgestalt/.
25. Zhang S, Know S, Snowdy J: WebGestalt: an integrated system for exploring gene sets in various biological contexts. Nuclic acids research 2009, 37(WEB Server issue):W541–W548.
26. Benjamin Y, Hochberg Y: Controlling the false discovery rate - a practical and powerful approach to multiple testing. J Roy Stat Soc B Met 1995, 57(1):289–300.
27. The Gene Ontology. http://www.genontology.org/.
28. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, et al: Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 2000, 25(1):25–29.
29. Kyoto Encyclopedia of Genes and Genomes. http://www.genome.jr/kegg/.
30. Kanekaha M, Goto S: KEGG: Kyoto Encyclopedia of Genes and Genomes. Nuclic Acids Res 2000, 28(1):27–30.
31. Nguyen CQ, Peck AB: Unraveling the pathophysiology of Sjogren syndrome-associated dry eye disease. Ocul Surf 2006, 4(1):11–27.
32. Mariette X, Gottenberg JE: Pathogenesis of Sjogren’s syndrome and therapeutic consequences. Curr Opin Rheumatol 2010, 22(5):471–477.
33. Sall K, Stevenson OD, Mundorf TK, Reis BL: Two multicenter, randomized studies of the efficacy and safety of cyclosporine ophthalmic emulsion in moderate to severe dry eye disease. CISA Phase 3 Study Group. Ophthalmology 2000, 107(4):631–639.
34. Ocholm P, Wintner K, Manthorpe R: Platelets in blood and salivary glands of patients with primary Sjogren’s Syndrome. Scand J Rheumatol Suppl 1986, 61:170–172.
35. Lodde EM, Sankar V, Kok MR, Leakan RA, Tak PP, Pillemner SR: Serum lipid levels in Sjögren’s syndrome. Rheumatology 2006, 45(4):481–484.
36. Ostrowski RA, Robinson JA: Antiphospholipid antibody syndrome and autoimmune diseases. Hematol Oncol Clin North Am 2008, 22(1):53–65.
37. Chai J, Logigian EL: Neurological manifestations of primary Sjogren’s syndrome. Curr Opin Neurol 2010, 23(5):509–513.
38. Caulfield VL, Balmer C, Dawson LJ, Smith PM: A role for nitric oxide-mediated glandular hypofunction in a non-apoptotic model for Sjogren’s syndrome. Rheumatology 2009, 48(7):727–733.

doi:10.1186/1471-2474-13-119
Cite this article as: Gorr et al: Text-mining applied to autoimmune disease research: the Sjögren’s syndrome knowledge base. BMC Musculoskeletal Disorders 2012 13:119.