Literature mining, gene-set enrichment and pathway analysis for target identification in Behçet's disease

P. Wilson[1], C. Larminie[1], R. Smith[2]

[1]Computational Biology, GlaxoSmithKline Medicine Research Centre, Herts, UK;
[2]Department of Renal Medicine, Addenbrooke's Hospital, Cambridge, UK.

Paul Wilson, MSc
Christopher Larminie, PhD
Rona Smith, MA, MB, BChir

Please address correspondence to:
Dr Rona Smith,
Box 57, Department of Renal Medicine,
Addenbrooke's Hospital,
Hills Road,
Cambridge CB2 0QQ, UK.
E-mail: ronasmith@doctors.org.uk

Received on June 13, 2016; accepted in revised form on September 7, 2016.

Clin Exp Rheumatol 2016; 34 (Suppl. 102): S101-S110.

Key words: Behçet's disease, text mining, gene-set enrichment, causal analysis

Competing interests: none declared.

ABSTRACT

Objective. *To use literature mining to catalogue Behçet's associated genes, and advanced computational methods to improve the understanding of the pathways and signalling mechanisms that lead to the typical clinical characteristics of Behçet's patients. To extend this technique to identify potential treatment targets for further experimental validation.*

Methods. *Text mining methods combined with gene enrichment tools, pathway analysis and causal analysis algorithms.*

Results. *This approach identified 247 human genes associated with Behçet's disease and the resulting disease map, comprising 644 nodes and 19220 edges, captured important details of the relationships between these genes and their associated pathways, as described in diverse data repositories. Pathway analysis has identified how Behçet's associated genes are likely to participate in innate and adaptive immune responses. Causal analysis algorithms have identified a number of potential therapeutic strategies for further investigation.*

Conclusion. *Computational methods have captured pertinent features of the prominent disease characteristics presented in Behçet's disease and have highlighted NOD2, ICOS and IL18 signalling as potential therapeutic strategies.*

Introduction

Behçet's disease (BD) (Disease Ontology DOID:13241) is a rare chronic relapsing inflammatory disorder of unknown aetiology characterised by recurrent oral and genital ulcerations with systemic manifestations, the most serious of which are ophthalmic, neurological, gastrointestinal and vascular in nature. BD is heterogeneous and lacks pathognomonic symptoms, laboratory or histological findings, and diagnosis relies on clinical criteria, such as, those of the International Study Group for Behçet's Disease (1). Generally, less complex mucocutaneous disease can be managed with topical steroids and non-steroidal agents, progressing to oral steroids and immunosuppressive drugs such as azathioprine, calcineurin inhibitors (cyclosporin and tacrolimus) or mycophenolate mofetil, in the minority of individuals who fail to respond to topical treatments, or whose disease progresses to involve other organs (2). More recently apremilast, an oral phosphodiesterase-4 inhibitor, has been shown to be effective treating oral ulcers in the short term, although whether this is sustained in the long term and the impact on other disease manifestations remains unclear (3). Anti-TNF agents are effective, but tend to be reserved for those with severe disease, or manifestations that are refractory to the above therapies. Interferon alpha, rituximab, alemtuzumab and cyclophosphamide are further therapeutic options for refractory disease (2). The pathogenesis of BD is likely to be multifactorial, with infectious agents and aberrations in B and T cell function being implicated on a background of genetic predisposing factors (4, 5).

In recent years there has been an "omics" revolution within medicine, following major advances in genetic sequencing techniques. Various computational approaches have proved critical to analyse the huge volume of data generated; to identify the genes, pathways and biological processes involved in the pathogenesis of several diseases, and subsequently prioritise possible drug targets (6). However, in a rare and heterogeneous condition, such as BD, the utility of such an approach is limited, particularly as transcriptomics data is influenced not only by the disease process itself, but also the immunosuppressive medications used to treat the disease.

That said, the available literature detailing observations on BD provides a huge wealth of experimental and clinical information, though the diversity of the condition and treatment approaches make systematically analysing the data in a gene-by-gene manner both difficult and time consuming. An alternative approach is to use a combination of literature mining techniques, where in essence, the available published literature (from a number of sources) is viewed as the equivalent of the experimental genomics or transcriptomics data. The resulting BD dataset is then analysed using various computational methodologies to identify the most likely regulation nodes that dictate gene activity, biological processes and disease pathways pertinent to BD (7).

This paper presents the data generated using such a literature data mining approach applied to BD. The most likely disease pathways are presented, and used as a hypothesis generating tool to identify areas of interest for further evaluation and potential novel therapeutic strategies.

Methods

Genes associated with BD were identified using three distinct text mining approaches. Previous experience has shown that reliance on one text mining strategy may fail to identify a significant numbers of associations reported in the scientific literature and that respective conclusions may not accurately reflect the complete knowledge base (8). The first search method retrieved gene identifiers from the NCBI (National Center for Biotechnology Information) Gene database (http://www.ncbi.nlm.nih.gov/gene/). The Gene database details gene-centric links to defined database fields, including literature, that include disease terms, and can be accessed via a simple text search interface that incorporates a number of easily implemented filters to restrict results. Behçet's-associated genes were identified by searching all fields for the term "Behçet*" and restricted to entries with the attributes "Human", "Current" (indicates annotation status) and "RefSeq" (a non-redundant sequence identifier).

A second approach used the Elsevier Text Mining Solution (https://www.elsevier.com/solutions/professional-services/text-mining) to identify human genes associated with BD. This is a licensed application and implements natural language processing (NLP) of full-text, unstructured literature (*i.e.* searches of abstract and discussion sections in addition to title, keywords etc.) to identify specific relationships including gene-disease relationships. This system is particularly powerful as search terms are expanded to include synonyms and abbreviations to identify co-occurring terms. Behçet's associated genes were identified using the terms "Behçet", "Genes", "Proteins" and "RNA" and restricted to MEDLINE and PubMed Central publications. An important feature of this approach is that sentences are extracted and co-occurring terms are highlighted which allows rapid evaluation and prioritisation of the results.

The third method utilised a Disease Maps approach (7, 9). A large network of human disease-associated genes that co-occur with Medical Subject Headings (MeSH) in Medline abstracts, more often than would be expected by chance, was constructed. The network was further extended to incorporate biological pathways and ontologies significantly enriched for the disease-associated genes such that the final entity comprised an extensive network of gene-to-disease terms-to-pathways. It has been reported that the emergent properties of such networks help identify novel disease relationships that would not be readily captured through conventional literature mining approaches (7, 9). A BD sub-network was then extracted from the global (*i.e.* the network that includes all gene-disease associations) human disease network.

Gene identifiers unique to only one of the search methods were manually validated to remove spurious associations, such as incorrect synonym extensions, to establish a high confidence dataset. The high confidence list of unique gene identifiers associated with BD compiled from the three search strategies was then used to complete gene enrichment analysis using the Enrichr application (10) and pathway analysis using Qiagen's Ingenuity Pathway®Analysis suite. The Enrichr web application (http://amp.pharm.mssm.edu/Enrichr) accepts a list of gene identifiers as input and computes gene enrichment scores for a comprehensive collection of gene-set libraries. Qiagen Ingenuity Pathway Analysis (IPA®) is a licensed web application used to further understand the most likely functional attributes of large 'omics' datasets. Both methods provide insights into the most likely collective biological functions of submitted gene lists by using statistical enrichment algorithms and prior biological knowledge derived from gene-set libraries (*e.g.* information extracted from peer reviewed studies, or the major biomedical repositories of pathway and molecular interaction data). The Ingenuity Upstream Regulator Analysis was also used to predict upstream regulators that may regulate gene expression of the Behçet's associated gene list. This algorithm is based on expected causal effects between known upstream regulators (*i.e.* findings derived from scientific literature and integrated within the Ingenuity® Knowledge Base) and lists of gene 'targets' (11).

Results

The three distinct text mining approaches (an NCBI Gene database search, the Elsevier Text Mining Solution and a Disease Map approach) identified 247 unique gene associations with BD (supplementary 1). The NCBI Gene search identified 176 gene associations, Elsevier text mining 96 and the Disease Map 92 gene identifiers. Commonality across the methods was assessed using a simple Venn analysis (Fig. 1). Only 29 gene identifiers were identified by all three of the methods and 93 identified by 2 or more of the methods. The distribution of the findings is not unexpected and merely reflects the individual strengths of each method; that is, the extensive NCBI curation effort, the implementation of Natural Language Processing by Elsevier, and statistically significant enrichment of co-occurring MeSH terms in Disease Map.

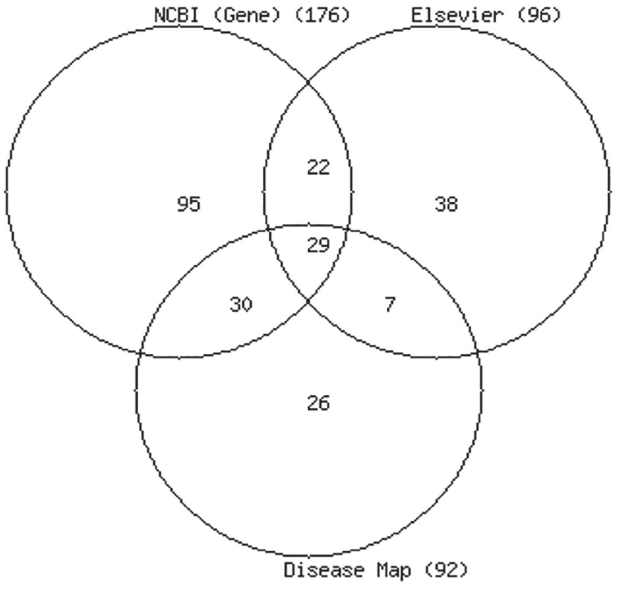

Fig. 1. Venn Diagram analysis comparing gene identifiers associated with Behçet's disease reported by NCBI, Elsevier and Disease Map text mining approaches. Numerical values indicate the number of unique gene identifiers included in each of the respective compartments.

Systematic gene-by-gene disease associations of the 247 unique gene identifiers retrieved from the Online Mendelian Inheritance in Man® (OMIM: http://omim.org/) highlighted a number of interesting functional associations relevant to a Behçet's phenotype. For example, several mutations in the NCSTN gene are causative of chronic skin disease characterised by recurrent scarring folliculitis (see http://omim.org/entry/605254). However, while a simple gene-by-gene search is of value, such an approach does not facilitate further insight as to how disease-associated genes are likely to interact, nor does it enable the emergent properties within a BD signalling network to be identified.

To better understand the functional significance of a genes list we must first associate the genes with their respective biological annotations, and then statistically determine if there is significant enrichment of known pathways and processes beyond what would be encountered in a random collection of gene identifiers, of similar size. We then infer that highly enriched (and statistically significant) features are indicative of the most likely underlying biological processes and that these are important starting points to investigate the most likely hierarchy and directionality of signalling events between the individual genes. Note that poorly connected (*i.e.* those genes that demonstrate no common ontology with other BD associated genes) are not reported as significant in such analyses.

The BD Disease Map was generated and then evaluated for credibility (*i.e.* did it capture the majority of reported BD disease features). This network was too large to easily visualise or manipulate in its entirety (comprising 644 nodes connected by 19220 edges) and was interactively investigated as a series of filtered sub-networks; one of which focussed on enrichment of Behçet's associated genes and eye disease terms (Fig. 2A). As anticipated uveitis and vasculitis were both identified as integral to this BD sub-network and incorporated a number of cytokines (including IL12B, IL17A and IL23A), HLA genes, TNF and TLR4, all of which have been previously reported as having significant roles in BD (12, 13). Also of note was inclusion of NOD2 as a highly connected node, with connecting edges to several genes and disease conditions, including sarcoidosis and ankylosing spondylitis. Mutations in this gene are reported to be causative in an autosomal dominant form of uveitis characteristic of Blau syndrome (14) and have been reported to be highly expressed in BAL cells retrieved from BD patients (15). However, several human studies evaluating three common NOD2 mutations as susceptibility genes for BD, found no significant association with BD and proposed that one of the mutations may be protective of BD (16, 17).

A second sub-network focussed on cell-types significantly associated with the MeSH disease term "Behçets" (Fig. 2B). This sub-network identified Th1, Th17, T-lymphocytes and neutrophils as important components of BD pathogenesis and provided details of Behçet's associated genes and their interactions within a Behçet's framework. In this sub-network several genes have edges connecting to multiple cell types and indicate important intra-cellular signalling events, for example, the ICOS gene node has edges to each of the T-cell nodes and has been associated with cytokine production in Th1 and Th17 cell types (18). The prominence of cytokine signalling and the imbalance in T-cell subsets in BD suggests that blockade of this co-stimulatory pathway may prove a tractable therapeutic target in the treatment of uveitis (19). These observations along with additional evaluation of sub-networks indicated that the disease map had captured many of the key features associated with BD and that the map could be used to explore additional, putative disease pathways.

To investigate how the 247 Behçet's associated genes identified may interact within a cellular signalling network and to identify potential therapeutic intervention strategies, gene-set enrichment and pathway analysis approaches were used. When the list of Behçet's associated genes was assessed relative to the Human Phenotype Ontology database (http://www.human-phenotype-ontology.org/) (20), using the Enrichr web application, the ranked results reported that the Behçet's literature defined list was significantly enriched for gene identifiers associated with several disease ontology terms, including arthralgia and vasculitis, that are frequent manifestations of BD (Table I). In each instance the standardised Z-score was close to or greater than two standard deviations below the mean and associated with highly significant adjusted p-values. Such scores are highly unlikely to be observed in a random or noisy (*i.e.* non-specific) gene list and strongly indicate that a number of the listed genes

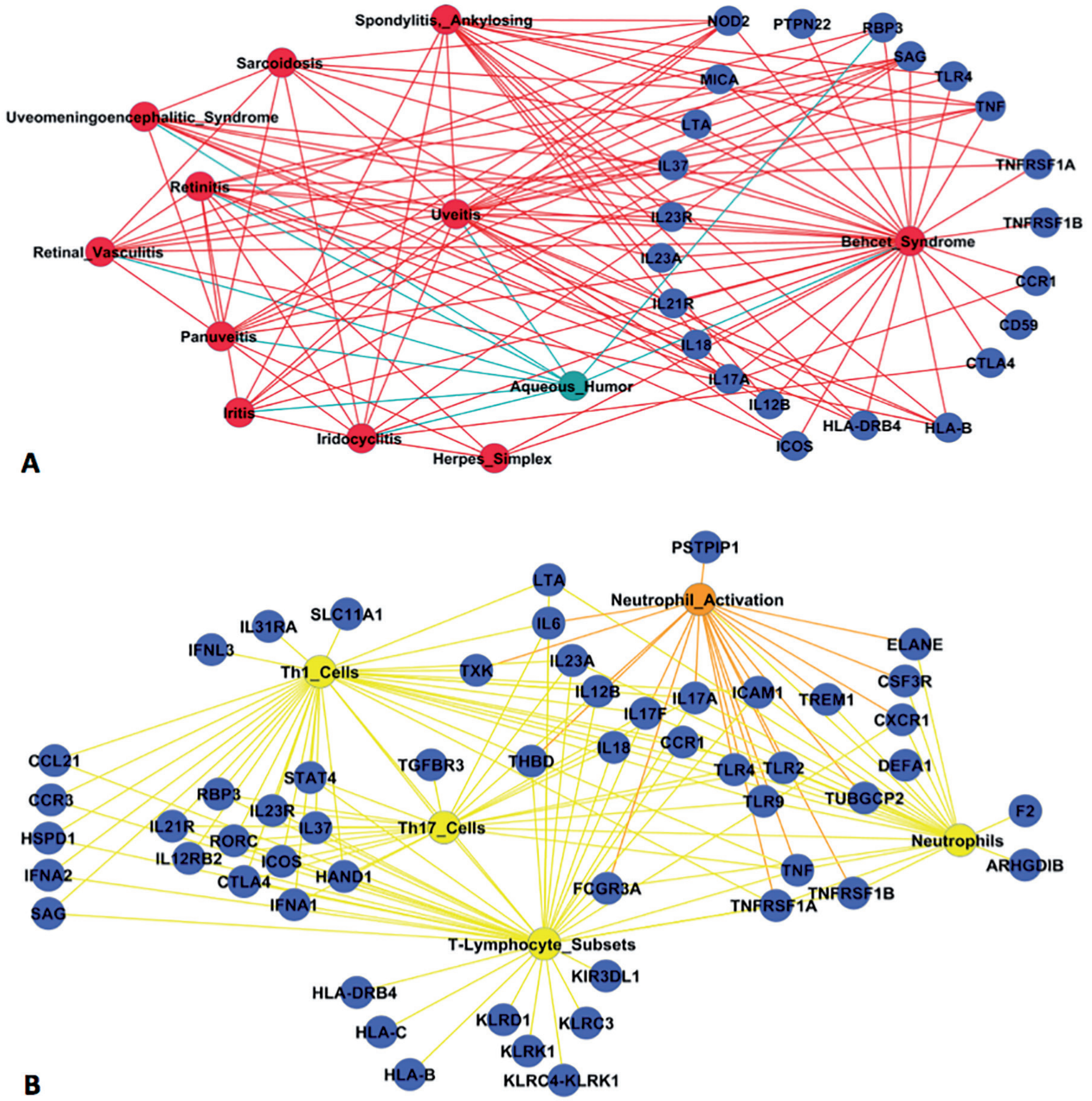

Fig. 2. A: A sub-network of the Behçet's disease network. Red nodes represent disease terms, yellow nodes represent cell types, blue represents gene nodes and orange nodes represent cellular processes. Edges represent publications with a statistically significant co-occurence of the two connected terms.
B: A sub-network detailing gene and immune cell interactions within a Behçet's disease framework. The sub-network helps better understand both gene-gene and gene-cell interactions and facilitates hypothesis generation in relation to gene and cell functional roles in Behçet's disease. The sub-network also indicates that neutrophil activation is a key contributor to the disease phenotype.

share common biological characteristics. The Enrichr application also lists the respective gene identifiers associated with each of the reported ontology terms for further detailed analysis (not shown). Observing common features of BD in the top ranks of the enrichment result again provided confidence that the Behçet's literature defined gene list captures many significant components of the underlying disease signalling mechanisms. Enrichr also completes enrichment analysis against many prior knowledge gene-set libraries and reports a comprehensive, system wide profiling that facilitates a deeper understanding of the most likely functional attributes 'encoded' within a query gene list. For example, it reported, via the Human gene Atlas (http://biogps.org/), that the 247 Behçet's associated gene list was significantly enriched (associated *p*-values of <0.001) for genes associated with CD55+ Natural Killer (NK) cells, CD33 Myeloid cells and CD14+ monocytes (not shown). While enrichment of WikiPathways 2015 (http://www.wikipathways.org/) reported significant enrichment (*p*-values <0.0001) of the human allograft rejec-

Table I. Summary of enriched disease terms identified in the merged Behçet's disease gene list. The summary statistics were generated via tables derived from the Human Phenotype Ontology database using the Enrichr web application. The Adjusted *p*-value column reports the significance associated with the Fisher's Exact Test applied of two gene sets. The Z-score is a correction of the Fisher's Exact Test standardised using background scores derived from random datasets, and the Combined Score is a product of the previous two values.

Term	Adjusted *p*-value	Z-score	Combined Score
Arthralgia (HP:0002829)	4E-14	-2.20	67.68
Gangrene (HP:0100758)	1.07E-10	-2.19	50.17
Pulmonary infiltrates (HP:0002113)	2.03E-10	-2.10	46.90
Pulmonary embolism (HP:0002204)	7.05E-10	-2.22	46.68
Abnormality of the pleura (HP:0002103)	2.03E-10	-1.98	44.23
Thrombophlebitis (HP:0004418)	2.03E-10	-1.98	44.19
Haemoptysis (HP:0002105)	8.94E-10	-2.08	43.28
Vasculitis (HP:0002633)	2.03E-10	-1.93	43.05
Abnormality of the pericardium (HP:0001697)	7.81E-10	-1.99	41.66
Arterial thrombosis (HP:0004420)	8.95E-10	-1.93	40.29
Myositis (HP:0100614)	2.03E-10	-1.80	40.20
Anorexia (HP:0002039)	1.84E-07	-2.18	33.82
Splenomegaly (HP:0001744)	1.84E-07	-2.15	33.36
Orchitis (HP:0100796)	8.67E-10	-1.53	31.92
Migraine (HP:0002076)	2.25E-07	-2.06	31.59
Skin ulcer (HP:0200042)	1.89E-07	-2.04	31.54

tion pathway, human toll-like receptor signalling, the human cytokine and inflammatory response, folate metabolism and vitamin B12 metabolism (not shown). Combining these observations into a systems view of the data both highlights the most probable functionally cooperating genes, underlying the respective biological processes, and facilitates identification of putative relationships between the different signalling pathways.

Using Qiagen Ingenuity Pathway Analysis, the Behçet's associated gene list was further evaluated to identify the most likely biological processes associated with the Behçet's literature defined gene collection. One section of the analysis reported is that of the enriched Canonical Pathways (Fig. 3). Each hand curated interaction pathway and its associated statistics are represented as a series ranked horizontal bars. These results are statistically highly significant and are much greater than would be observed in a typical transcriptomics study (personal observation) and indicate that the BD gene list is highly enriched for genes that have been observed to cooperate in important biological processes. The ranked results indicate significant enrichment of genes known to play functional roles in cross-talk between innate and adaptive immune cells, interactions between dendritic cells and NK cells, interactions between macrophages, fibroblasts and endothelial cells, each of which may co-occur with alterations in T-cell signalling and differentiation, cytokine signalling and TREM1 signalling.

Each pathway was systematically further mined to understand how the associated genes are known to interact and prior knowledge evaluated within a BD context (not shown). Many of these investigations identified multiple densely connected cytokine and cytokine complexes as potential signalling hubs. However, both the number and high degree of connectivity of the underlying prior knowledge were too complex to understand in terms of either a hierarchical or simple directional effect, and suggest that complex feedback loops are used to regulate *in-situ* signalling events. To better understand the myriad of potential signalling mechanisms we applied the Ingenuity Upstream Regulator Analysis function to represent the BD derived gene collection in terms of the most likely upstream regulators (*i.e.* those genes that have been previously reported to affect the expression of another gene). This reported that the majority of genes in the BD list could be regulated by complex polysaccharides and a number of cytokines. These included IFN gamma, TNF, IL10, IL-1Beta, IL2, IL6 and the IL12 complex (Table II). Again, the statistical significance associated with the respective predictions is of a greater magnitude of confidence than that generally observed in a typical transcriptomics study, and this indicates that the gene lists are significantly enriched for the component units of previously observed biological signalling networks. When the BD gene list was analysed using Ingenuity's Causal Network Analysis algorithm (which identifies causal connections between diseases, genes and networks of upstream regulators), and the knowledge space restricted to human experimentally validated findings, IL18 was identified as the most plausible master upstream regulator (*i.e.* the regulator that best explains connectivity between the observed upstream regulators reported in Table II). A further variation on the predictive upstream function is to identify drug biologics and small molecules that may regulate a given gene list (not shown). This approach identified alefacept (an anti-CD2 molecule), etanercept (an anti-TNF-α), infliximab (an anti-TNF-α) and fontolizumab (an anti-IFN gamma) as biologics most likely to interact with many of genes in the Behçet's associated gene list.

Discussion

BD is a complex auto-inflammatory syndrome with diverse clinical manifestations. Management of BD remains challenging for many reasons; difficulties making an initial diagnosis and evaluating disease activity; heterogeneity in disease severity and organ manifestations, and variability in disease response to treatment, both between patients, and even different manifestations within an individual.

A huge body of scientific literature on BD exists. However, even with stringent filtering, PubMed returns more than 1000 published papers, detailing a multitude of observations describing gene focussed studies through to several genome wide association studies, which have identified several susceptibility loci, that include IL10, IL23R-

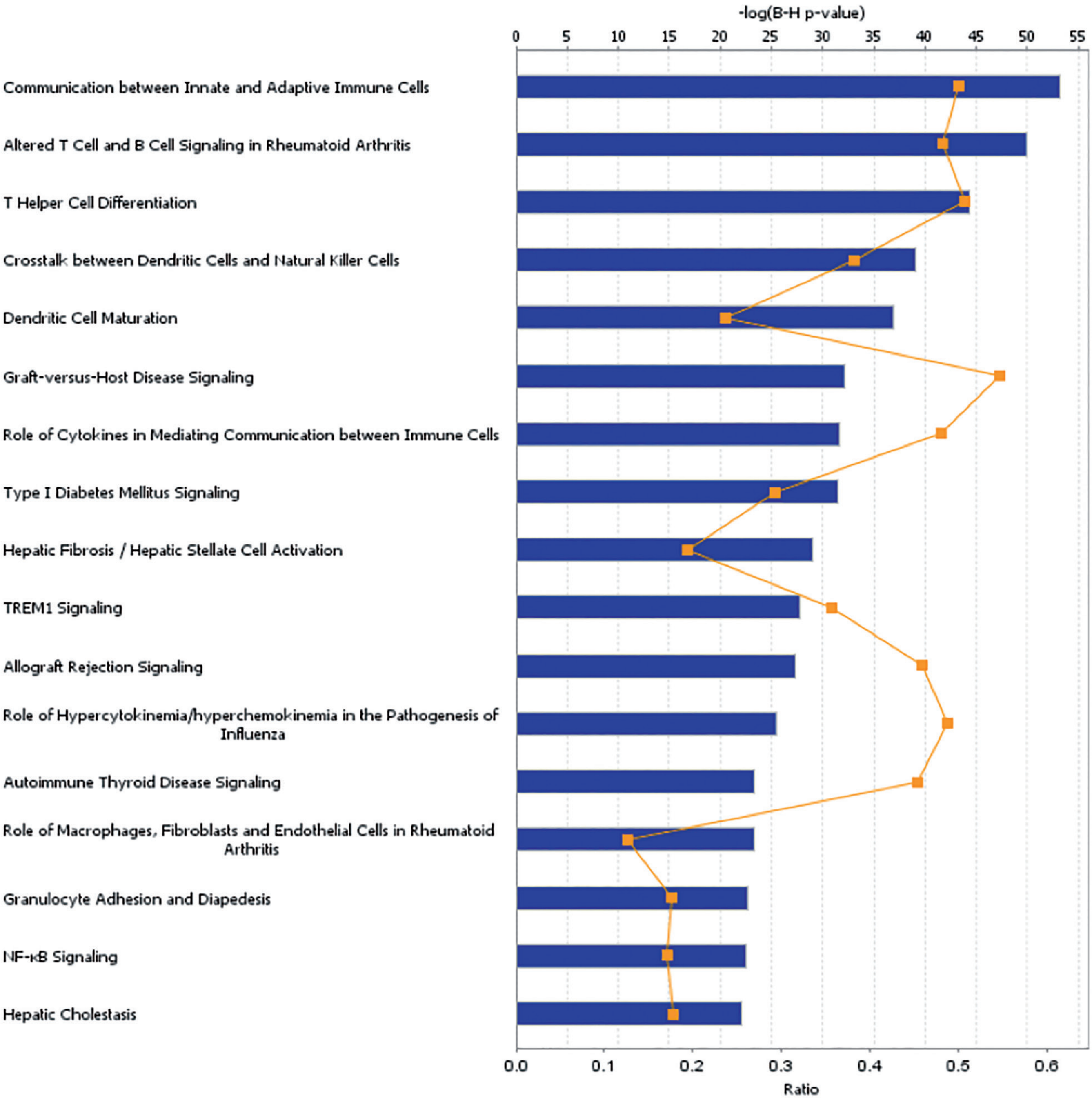

Fig. 3. Graphical summary of the Qiagen Ingenuity Canonical Pathways reported as significantly enriched when the merged list of Behçet's disease associated gene identifiers are assessed. The magnitude of each bar indicates the statistical significance of enrichment between two gene-sets as determined by the Benjamin-Hochberg Multiple Testing Correction *p*-value. Ingenuity Canonical pathways can also be ordered by the ratio value, which is the number of molecules in a given pathway that meet the specified cut-off criteria, divided by total number of molecules that make up that pathway (indicated by the orange trace line). The high scores and high ratios associated with the enrichment profile indicates significant perturbation of the immune system in Behçet's disease and that the disease effect includes effects on T-cell differentiation and cytokine signalling between several types of immune cells.

IL12RB2 and STAT4 (21-25). One approach is to investigate each of these associations in a gene-by-gene manner or even extend this to a pathway understanding. However, given the genetic complexity and diverse disease manifestations this may prove to be of limited effect as it will most likely fail to effectively integrate and contextualise such a large volume of information into a 'global' decision making knowledgebase. We propose that applying text mining techniques in combination with gene-enrichment and pathway analysis tools will more readily facilitate integration of this huge compendium into a manageable Behçet's focussed framework. Such disease network repositories should enable clinicians to better understand the most likely molecular interactions underlying disease characteristics and will also be amenable to causal analysis tools to elucidate the regulatory mechanism driving disease

Table II. The IPA Upstream Regulator Analysis tool annotates a gene as an upstream regulator if it has been reported to affect the expression of another gene, and uses a computational algorithm to predict if these upstream regulators are likely to regulate large numbers of genes in a gene list. The predicted top ten regulators of the merged list of Behçet's-associated gene network are shown. The comparative highly significant statistics indicate that the regulators are likely to 'cross-talk' to define disease features and that it is unlikely that there is one prominent driver. The prediction was generated using data derived from human experimental data.

Upstream regulator	Target molecules	*p*-value of over lap
lipopolysaccharide	142	8,93E-101
IFNG	121	7,47E-90
TNF	121	2,07E-77
IL10	71	4,51E-73
IL1B	93	2,74E-72
IL2	78	1,90E-71
IL6	69	3,21E-64
poly rI:rC-RNA	69	3,44E-63
TGFB1	107	1,16E-60
IL4	52	3,51E-59

signalling networks and ideally, ultimately support the selection of novel therapeutic strategies.

To implement such a strategy gene identifiers associated with the disease term Behçets were retrieved using three independent literature mining tools. One of these defined an interactive BD Map comprising 644 nodes connected by 19220 edges. This data structure provided a rich resource detailing how the Behçets-associated genes are known to interact and that these genes were significantly enriched both for many disease mesh terms characteristic of Behçets disease, and for biological pathways and cell types reported to contribute to inflammatory disease conditions. It is important to appreciate that the identified genes represent disease associations extracted from PubMed and that there is no direct quantification of the validity of the reported associations and furthermore, that a gene may be associated with contradictory findings. However, the strength of the method is that we are evaluating a gene collective that is reported to determine the BD phenotype, and that the relevant underlying biology connecting the gene-set will be significantly enriched. It has been our experience that the majority of spurious reports will be omitted from the network as they will be poorly connected and therefore statistically insignificant relative to the main body of the dataset.

The Behçets disease network identified NOD2 as an important hub gene in a subnetwork focussing on eye disease. The edges associated with the NOD2 node detail pivotal roles in innate immunity and autoinflammation (27), bacterial sensing (14), the regulation of innate immunity (28), and synergistic effects with the TLRs to induce pro-inflammatory cytokine responses (15). This suggests that a deeper understanding of NOD2 signalling events may prove an important focus of future therapeutic intervention. Of interest is that genetic studies of NOD2 have not validated this gene as a susceptibility locus for BD (17, 25) and that a common Crohn's associated mutation may be protective in a Caucasian population (16). Such observations indicate that NOD2 is more likely to direct a disease effect via synergistic interactions with other gene products, such as the TLRs and TNFAIP3 (15, 29). The ICOS node also appeared as an important hub in several sub-network analyses and the associated edges provided details of important roles in Th1 and Th17 cytokine production (18), and autoimmune disease (30). Furthermore, it has been reported that ICOS expression is a potential marker of disease activity (19) and has been demonstrated to be amenable to immunomodulation treatment in animal models of autoimmune disease (31).

The Behçets disease map was also used to investigate relationships between Behçet's associated genes and immune cell types. This analysis indicates a complex interplay between neutrophils and T-lymphocytes, and details several putative signalling mechanisms that inter-connect the innate and adaptive immune mechanisms. Of particular interest, it has been reported that a unique subset of CXCL8/GM-CSF-producing T cells, that differ from both Th1 and Th2 T cells, appear to promote neutrophil-rich pathologies of chronic autoinflammatory disease (32, 33). We propose that a detailed exploration of how these inflammatory signalling mechanisms inter-regulate in a disease scenario is likely to facilitate novel therapeutic intervention strategies.

To generate a super resource that maximised the volume of Behçet's associated observations from the published corpus we integrated findings using three independent text mining strategies. The combined methods identified 247 Behçet's associated genes, and this compendium was used to further elucidate the most likely signalling pathways and regulatory mechanisms that operate within a Behçet's phenotype. Gene-set enrichment analysis indicated that this more comprehensive BD gene compendium captured many of the pertinent characteristics of BD, and the ten most significant ontology terms reported identified arthralgia and vasculitis as prominent terms, both of which are considered key features of BD (1, 34). Both the prominence and statistical significance of these terms added to our confidence that the BD gene list was highly enriched (relative to that observed in a random gene-set of equivalent size) for genes that determine clinical manifestations of BD. Again, we anticipate that 'spurious' gene-disease associations are unlikely to be enriched with the same ontological terms and will not contribute to the statistically significant terms. These ontology-associated terms are themselves amenable to a further iteration of focussed analysis. For example, overlaying known drug interactions of the gene identifiers that defined the arthralgia hit (*i.e.* IL1RN, IL23R, PTPN22, NOD2, MEFV, C4A, PSTPIP1, STAT4, CTLA4, NLRP3, IL12B, CCR6, ICOS, JAK2, IL12RB2, IL10, MMP2, HLA-B, MIF, TNFRSF1A, IL6, IL2RA, IRF5, TLR4, PTPN2 AND HLA-DRB1) indicate a putative therapeutic role for the TLR4 inhibitor Resatorvid (35) (not shown).

A notable ontology term that is absent is uveitis. Further investigation of the Human Phenotype Ontology (HPO) database confirmed that the uveitis disease ontology is currently under-populated (see http://www.human-phenotype-ontology.org/hpoweb/showterm?id=HP:0000554) and would not generate a statistically significant hit. It should be noted that HPO curators welcome input from those with expert domain knowledge to help improve the human phenotype ontologies.

Gene-set enrichment also identified several immune cell types and putative co-factors as likely mediators of a BD phenotype. This included high scoring hits for CD55$^+$ NK cells, CD33 myeloid and CD14$^+$ monocytes. Further literature mining readily identified supporting literature detailing a possible role for monocytes in the orchestration of bacterial-induced innate immune responses (36), and a dysfunctional regulation of the chemokine CXCL10 (37). Zinc deficiency has been reported in patients presenting with recurrent apthous stomatitis (RAS) (38) and folate supplementation has been proposed as a treatment of RAS (39). Combining these divergent observations encapsulates many BD characteristics, which when combined should help direct future therapeutic intervention strategies.

As one would anticipate with an auto-inflammatory condition, Canonical Pathway analysis identified significant effects on T-cell, B-cell, dendritic and macrophage cell types. As with other enrichment methods the underlying genes can be retrieved and investigated for gene-gene interactions, upstream regulator predictions and for potential cross-talk events between pathways (not shown). Of note was several pathways attributed to macrophage signalling, including; “Role of Macrophages, Fibroblasts and Endothelial Cells in Rheumatoid Arthritis”, “Differential Regulation of Cytokine Production in Macrophages and T Helper Cells by IL17A and IL17F and IL12 Signalling and Production in Macrophages” (not shown). It is of interest that neither the macrophage nor monocyte cell types were identified in our BD Disease Network. A follow-up of analysis indicated that a hit for the Mesh term ‘macrophages’ did not pass a threshold for statistical significance and was therefore not included in the resulting Disease Network. However, our BD Network did identify the MCP-1 (CCL2) gene as a highly connected gene node and the MCP-1 gene node was readily identified as a highly connected hub in several sub-networks (not shown). This cytokine is chemotactic for monocytes and has been implicated in the pathogenesis of diseases characterised by monocytic infiltrates (40) and associated with clinical severity in Behçet’s disease (41). Combined these indicated an important role for macrophage cell types in the BD phenotype. This emergent property of the network is of particular interest as therapeutic agents targeting MCP-1 may offer alternative therapeutic strategies in treatment of BD.

Predicting the most likely Upstream Regulators of the BD gene list identified a number of polysaccharides and cytokines as the most likely ‘drivers’ of the BD gene collection. The former is supportive of the belief that infectious agents are important in the development of BD (42), however, it may also be indicative of cellular damage due to over-active immune response. The predicted cytokine regulators include IL12. This is of particular interest in a BD context as macrophages are reported to produce IL12 when they encounter a pathogen (43). A major effect of IL12 signalling on macrophages is the induction of interferon gamma which favours the differentiation of T helper 1 cells that regulate the adaptive immune response. Aberrations in both B and T cell function is a well characterised feature of BD so further understanding of how signalling between these cytokines is perturbed in a BD phenotype may help direct future therapeutic intervention strategies. Further pathway analysis of the predicted cytokine upstream regulators generated a highly interconnected module suggestive of complex feedback loops (not shown). The high degree of complexity indicates that targeting any one cytokine is unlikely to have long-lasting therapeutic benefits and that a pan-cytokine inhibitor may prove to be a more valid treatment strategy.

The large number of connections, within the Behçet’s associated genes, complicated both our understanding of potential disease signalling mechanisms and weakened our ability to generate therapeutic hypotheses. To address both issues we restricted the available knowledge space to include only human experimentally validated findings. Ingenuity’s Causal Network Analysis algorithm then identified IL18 as the most plausible master upstream regulator. This finding is of interest as it provides a model of the dataset that connects a substantial number of the Behçet’s associated genes using only human experimental data, and provides a readily testable hypothesis. Interestingly, several publications report IL18 gene polymorphisms associations with BD in human cohorts (44, 45). Furthermore, serum levels of IL18 were observed to be significantly higher in patient subgroups relative to healthy controls and found to be correlated with the disease activity score (46, 47). In addition, Th1 cells expressing both TXK and IL18 (along with other Th1-associated cytokines) may be causative in the development of both skin and intestinal lesions in Behçet’s patients (48, 49). These combined observations suggest that IL18 is an important mediator of the Th1 cytokine response and deserves further scrutiny as a potential therapeutic target.

BD is classified as a rare disease (ORPHA117, http://www.orpha.net/), however, despite a minor contribution to overall global disease the volume of associated literature is too large to digest in a practical manner. We have illustrated that combining text mining methodologies enable comprehensive integration of a large volume of gene-disease associated information that is amenable to pathway enrichment analysis. We have also exemplified how computational methods may be used to better understand both the connectivity and the directionality of signalling events that are most likely relating the diverse manifestations of BD. We propose that such methods will identify key regulatory nodes and that this knowledge will contribute to the formulation of novel treatment strategies.

Acknowledgements

We would like to thank Caroline Savage and Philippe Sanseau for their support and comments on the manuscript, and the helpful suggestions from the anonymous reviewers.

Supplementary data

Supplementary data are available at Clinical and Experimental Rheumatology Online.

References

1. ISGBD Criteria for diagnosis of Behçet's disease. International Study Group for Behçet's Disease. *Lancet* 1990; 335: 1078-80.
2. BEHÇET'S DISEASE: Drug pathway, 2013 http://www.Behçets.nhs.uk/images/downloads/Bechets%20drug%20pathway%20v20-%20July%2025%202013.pdf
3. HATEMI G, MELIKOGLU M, TUNC R *et al.*: Apremilast for Behçet's syndrome--a phase 2, placebo-controlled study. *N Engl J Med* 2015; 372: 1510-8.
4. DIRESKENELI H: Behçet's disease: infectious aetiology, new autoantigens, and HLA-B51. *Ann Rheum Dis* 2001; 60: 996-1002.
5. DE MENTHON M, LAVALLEY MP, MALDINI C, GUILLEVIN L, MAHR A: HLA-B51/B5 and the risk of Behçet's disease: a systematic review and meta-analysis of case-control genetic association studies. *Arthritis Rheum* 2009; 61: 1287-96.
6. SCHADT EE, BJÖRKEGREN JL: NEW: network-enabled wisdom in biology, medicine, and health care. *Sci Transl Med* 2012; 4: 115rv1.
7. LI Y, AGARWAL P: A pathway-based view of human diseases and disease relationships. *PLoS One* 2009; 4: e4346.
8. SAMADZADEH GR, RIGI T, GANJALI AR: Comparison of Four Search Engines and their efficacy With Emphasis on Literature Research in Addiction (Prevention and Treatment). *Int J High Risk Behav Addict* 2012; 1: 166-71.
9. CHAN SY, LOSCALZO J: The emerging paradigm of network medicine in the study of human disease. *Circ Res* 2012; 111: 359-74.
10. CHEN EY, TAN CM, KOU Y *et al.*: Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinformatics* 2013; 14: 128.
11. KRÄMER A, GREEN J, POLLARD J JR, TUGENDREICH S: Causal analysis approaches in Ingenuity Pathway Analysis. *Bioinformatics* 2014; 30: 523-30.
12. ZHOU ZY, CHEN SL, SHEN N, LU Y: Cytokines and Behçet's disease. *Autoimmun Rev* 2012; 10: 699-704.
13. DO JE, KWON SY, PARK S, LEE ES: Effects of vitamin D on expression of Toll-like receptors of monocytes from patients with Behçet's disease. *Rheumatology* (Oxford) 2008; 47: 840-8.
14. ROSENBAUM JT, ROSENZWEIG HL, SMITH JR, MARTIN TM, PLANCK SR: Uveitis secondary to bacterial products. *Ophthalmic Res* 2008; 40: 165-8.
15. HAMZAOUI K, ABID H, BERRAIES A, AMMAR J, HAMZAOUI A: NOD2 is highly expressed in Behçet disease with pulmonary manifestations. *J Inflamm* 2012; 9: 3.
16. KAPPEN JH, WALLACE GR, STOLK L *et al.*: Low prevalence of NOD2 SNPs in Behçet's disease suggests protective association in Caucasians. *Rheumatology* 2009; 48: 1375-7.
17. UYAR FA, SARUHAN-DIRESKENELI G, GÜL A: Common Crohn's disease-predisposing variants of the CARD15/NOD2 gene are not associated with Behçet's disease in Turkey. *Clin Exp Rheumatol* 2004; 22 (Suppl. 34): S50-2.
18. USUI Y: Expression and function of ICOS on CD4 T cells and application to therapy in patients with ocular Behçet's disease with uveitis. *Nihon Ganka Gakkai Zasshi* 2012; 116: 1037-45.
19. USUI Y, TAKEUCHI M, YAMAKAWA N *et al.*: Expression and function of inducible costimulator on peripheral blood CD4$^+$ T cells in Behçet's patients with uveitis: a new activity marker? *Invest Ophthalmol Vis Sci* 2010; 51: 5099-104.
20. KÖHLER S, DOELKEN SC, MUNGALL CJ *et al.*: The Human Phenotype Ontology project: linking molecular biology and disease through phenotype data. *Nucleic Acids Res* 2014; 42: D966-74.
21. REMMERS EF, COSAN F, KIRINO Y *et al.*: Genome-wide association study identifies variants in the MHC class I, IL10, and IL23R-IL12RB2 regions associated with Behçet's disease. *Nat Genet* 2010; 42: 698-702.
22. MIZUKI N, MEGURO A, OTA M *et al.*: Genome-wide association studies identify IL23R-IL12RB2 and IL10 as Behçet's disease susceptibility loci. *Nat Genet* 2010; 42: 703-6.
23. XAVIER JM, SHAHRAM F, DAVATCHI F *et al.*: Association study of IL10 and IL23R-IL12RB2 in Iranian patients with Behçet's disease. *Arthritis Rheum* 2012; 64: 2761-72.
24. HOU S, KIJLSTRA A, YANG P: The genetics of Behçet's disease in a Chinese population. *Front Med* 2012; 6: 354-9.
25. KIRINO Y, BERTSIAS G, ISHIGATSUBO Y *et al.*: Genome-wide association analysis identifies new susceptibility loci for Behçet's disease and epistasis between HLA-B*51 and ERAP1. *Nat Genet* 2013; 45: 202-7.
26. SHIGEMURA T, KANEKO N, KOBAYASHI N *et al.*: Novel heterozygous C243Y A20/TNFAIP3 gene mutation is responsible for chronic inflammation in autosomal-dominant Behçet's disease. *RMD Open* 2016; 2: e000223.
27. BORZUTZKY A, FRIED A, CHOU J, BONILLA FA, KIM S, DEDEOGLU F: NOD2-associated diseases: Bridging innate immunity and autoinflammation. *Clin Immunol* 2010; 134: 251-61.
28. CASO F, COSTA L, RIGANTE D *et al.*: Caveats and truths in genetic, clinical, autoimmune and autoinflammatory issues in Blau syndrome and early onset sarcoidosis. *Autoimmun Rev* 2014; 13: 1220-9.
29. SHIGEMURA T, KANEKO N, KOBAYASHI N *et al.*: Novel heterozygous C243Y A20/TNFAIP3 gene mutation is responsible for chronic inflammation in autosomal-dominant Behçet's disease. *RMD Open* 2016; 2:e000223.
30. DONG C, NURIEVA RI: Regulation of immune and autoimmune responses by ICOS. *J Autoimmun* 2003; 21: 255-60.
31. FREY O, MEISEL J, HUTLOFF A *et al.*: Inducible costimulator (ICOS) blockade inhibits accumulation of polyfunctional T helper 1/T helper 17 cells and mitigates autoimmune arthritis. *Ann Rheum Dis* 2010; 69: 1495-501.
32. SCHAERLI P, BRITSCHGI M, KELLER M *et al.*: Characterization of human T cells that regulate neutrophilic skin inflammation. *J Immunol* 2004; 173: 2151-8.
33. KELLER M, SPANOU Z, SCHAERLI P *et al.*: T cell-regulated neutrophilic inflammation in autoinflammatory diseases. *J Immunol* 2005; 175: 7678-86.
34. MAT MC, SEVIM A, FRESKO I, TÜZÜN Y: Behçet's disease as a systemic disease. *Clin Dermatol* 2014; 32: 435-42.
35. MATSUNAGA N, TSUCHIMORI N, MATSUMOTO T, LI M: TAK-242 (resatorvid), a small-molecule inhibitor of Toll-like receptor (TLR) 4 signaling, binds selectively to TLR4 and interferes with interactions between TLR4 and its adaptor molecules. *Mol Pharmacol* 2011; 79: 34-41.
36. EKŞIOGLU-DEMIRALP E, KIBAROGLU A, DIRESKENELI H *et al.*: Phenotypic characteristics of B cells in Behçet's disease: increased activity in B cell subsets. *J Rheumatol* 1999; 26: 826-32.
37. AMBROSE N, KHAN E, RAVINDRAN R *et al.*: The exaggerated inflammatory response in Behçet's syndrome: identification of dysfunctional post-transcriptional regulation of the IFN-γ/CXCL10 IP-10 pathway. *Clin Exp Immunol* 2015; 181: 427-33.
38. OZLER GS: Zinc deficiency in patients with recurrent aphthous stomatitis: a pilot study. *J Laryngol Otol* 2014; 128: 531-3.
39. KALKAN G, KARAKUS N, YIGIT S: Association of MTHFR gene C677T mutation with recurrent aphthous stomatitis and number of oral ulcers. *Clin Oral Investig* 2014; 18: 437-41.
40. MACDERMOTT RP, SANDERSON IR, REINECKER HC: The central role of chemokines (chemotactic cytokines) in the immunopathogenesis of ulcerative colitis and Crohn's disease. *Inflamm Bowel Dis* 1998; 4: 54-67.
41. KIM SK, JANG WC, AHN YC *et al.*: Promoter -2518 single nucleotide polymorphism of monocyte chemoattractant protein-1 is associated with clinical severity in Behçet's disease. *Inflamm Res* 2012; 61: 541-5.

42. LEHNER T: The role of heat shock protein, microbial and autoimmune agents in the aetiology of Behçet's disease. *Int Rev Immunol* 1997; 14: 21-32.
43. TRINCHIERI G: Cytokines acting on or secreted by macrophages during intracellular infection (IL-10, IL-12, IFN-gamma). *Curr Opin Immunol* 1997; 9: 17-23.
44. XU Y, ZHOU K, YANG Z *et al.*: Association of cytokine gene polymorphisms (IL-6, IL-12B, IL-18) with Behçet's disease : A meta-analysis. *Z Rheumatol* 2016 Jan 22 [Epub ahead of print].
45. LEE YJ, KANG SW, PARK JJ *et al.*: Interleukin-18 promoter polymorphisms in patients with Behçet's disease. *Hum Immunol* 2006; 67: 812-8
46. MUSABAK U, PAY S, ERDEM H *et al.*: Serum interleukin-18 levels in patients with Behçet's disease. Is its expression associated with disease activity or clinical presentations? *Rheumatol Int* 2006; 26: 545-50.
47. OZTAS MO, ONDER M, GURER MA, BUKAN N, SANCAK B: Serum interleukin 18 and tumour necrosis factor-alpha levels are increased in Behçet's disease. *Clin Exp Dermatol* 2005; 30: 61-3.
48. NAGAFUCHI H, TAKENO M, YOSHIKAWA H *et al.*: Excessive expression of Txk, a member of the Tec family of tyrosine kinases, contributes to excessive Th1 cytokine production by T lymphocytes in patients with Behçet's disease. *Clin Exp Immunol* 2005; 139: 363-70.
49. SUZUKI N, NARA K, SUZUKI T: Skewed Th1 responses caused by excessive expression of Txk, a member of the Tec family of tyrosine kinases, in patients with Behçet's disease. *Clin Med Res* 2006; 4: 147-51.