Journal of proteome research

Article

Cite This: J. Proteome Res. 2019, 18, 4231−4239

pubs.acs.org/jpr

Mining the Proteome Associated with Rheumatic and Autoimmune Diseases

Cristina Ruiz-Romero,[*,†] Maggie P. Y. Lam,[‡] Peter Nilsson,[§] Patrik Önnerfjord,[∥] Paul J. Utz,[⊥] Jennifer E. Van Eyk,[#] Vidya Venkatraman,[#] Justyna Fert-Bober,[#] Fiona E. Watt,[▽] and Francisco J. Blanco[*,○]

[†] Grupo de Investigación de Reumatología (GIR), Unidad de Proteómica, INIBIC − Complejo Hospitalario Universitario de A Coruña, SERGAS, Universidad de A Coruña, A Coruña 15006, Spain

[‡] Department of Medicine, Division of Cardiology, Consortium for Fibrosis Research and Translation, Anschutz Medical Campus, University of Colorado Denver, Aurora, Colorado 80045, United States

[§] Division of Affinity Proteomics, SciLifeLab, Department of Protein Science, KTH Royal Institute of Technology, Stockholm 17121, Sweden

[∥] Department of Clinical Sciences, Section for Rheumatology and Molecular Skeletal Biology, Lund University, Lund 22184, Sweden

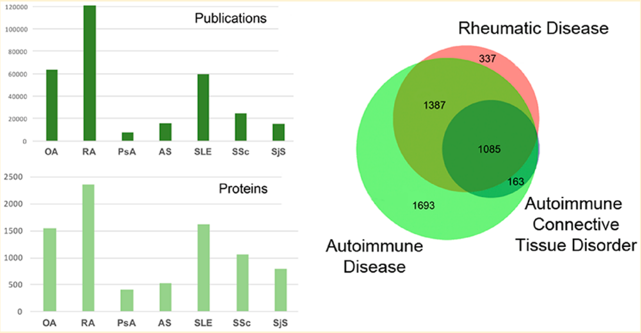
[⊥] Division of Immunology and Rheumatology, Stanford University School of Medicine; Palo Alto, California 94304, United States

[#] Department of Medicine and The Heart Institute, Cedars-Sinai Medical Center, Los Angeles, California 90048, United States

[▽] Arthritis Research UK Centre for Osteoarthritis Pathogenesis, Kennedy Institute of Rheumatology, University of Oxford, Oxford OX3 7FY, United Kingdom

[○] Grupo de Investigación de Reumatología, INIBIC-Complejo Hospitalario Universitario de A Coruña, SERGAS, Departamento de Medicina Universidad de A Coruña, A Coruña 15006, Spain

Ⓢ *Supporting Information*



ABSTRACT: A steady increase in the incidence of osteoarthritis and other rheumatic diseases has been observed in recent decades, including autoimmune conditions such as rheumatoid arthritis, spondyloarthropathies, systemic lupus erythematosus, systemic sclerosis, and Sjögren's syndrome. Rheumatic and autoimmune diseases (RADs) are characterized by the inflammation of joints, muscles, or other connective tissues. In addition to often experiencing debilitating mobility and pain, RAD patients are also at a higher risk of suffering comorbidities such as cardiovascular or infectious events. Given the socioeconomic impact of RADs, broad research efforts have been dedicated to these diseases worldwide. In the present work, we applied literature mining platforms to identify "popular" proteins closely related to RADs. The platform is based on publicly available literature. The results not only will enable the systematic prioritization of candidates to perform targeted proteomics studies but also may lead to a greater insight into the key pathogenic processes of these disorders.

KEYWORDS: *Human Proteome Project, rheumatic diseases, autoimmune diseases, osteoarthritis, bioinformatics*

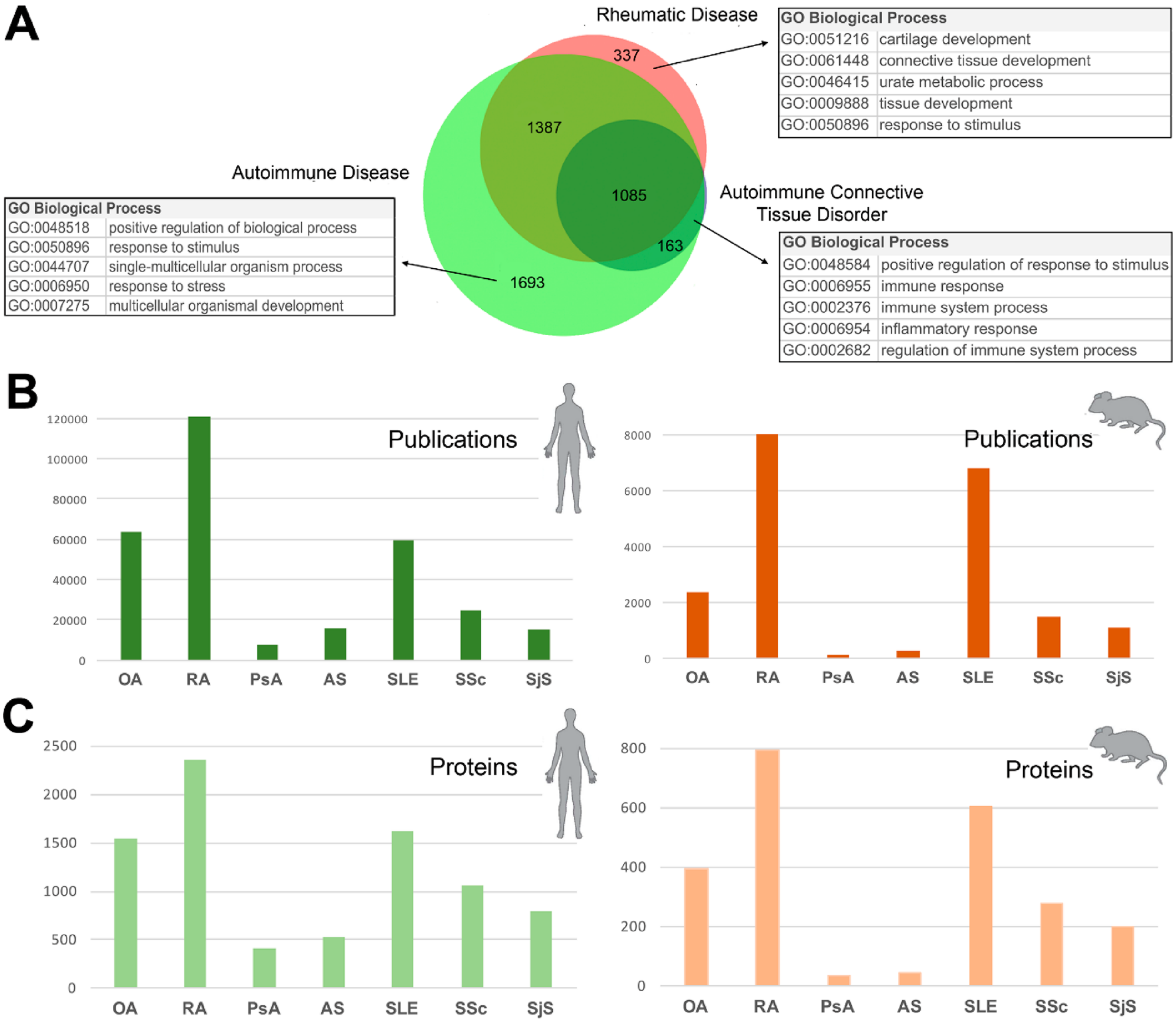
1. INTRODUCTION

Rheumatic and musculoskeletal diseases (RMDs) are pathological conditions affecting joints and connective tissues, causing intermittent or chronic pain and inflammation. This term covers more than 200 different conditions whose etiologies, epidemiologies, and clinical manifestations can differ widely,[1] such as osteoarthritis (OA), rheumatoid arthritis (RA), or systemic lupus erythematosus (SLE). However, they are globally characterized by their overall high prevalence in the general population, their tendency for chronicity, their potential to cause disability or functional limitation, and, in many cases, an association with increased cardiovascular mortality and morbidity.[2] The results of the Global Burden of Disease study carried out by the World Health Organization (WHO) in 2010 showed that the prevalence and morbidity of these diseases is exceptionally high worldwide, accounting for 21.3% of the total global years lived with disability.[3] For all of these reasons, RMDs have a great impact on the quality of life

Special Issue: Human Proteome Project 2019

Received: May 31, 2019
Published: October 10, 2019

DOI: 10.1021/acs.jproteome.9b00360



GO Biological Process	
GO:0051216	cartilage development
GO:0061448	connective tissue development
GO:0046415	urate metabolic process
GO:0009888	tissue development
GO:0050896	response to stimulus

GO Biological Process	
GO:0048518	positive regulation of biological process
GO:0050896	response to stimulus
GO:0044707	single-multicellular organism process
GO:0006950	response to stress
GO:0007275	multicellular organismal development

GO Biological Process	
GO:0048584	positive regulation of response to stimulus
GO:0006955	immune response
GO:0002376	immune system process
GO:0006954	inflammatory response
GO:0002682	regulation of immune system process

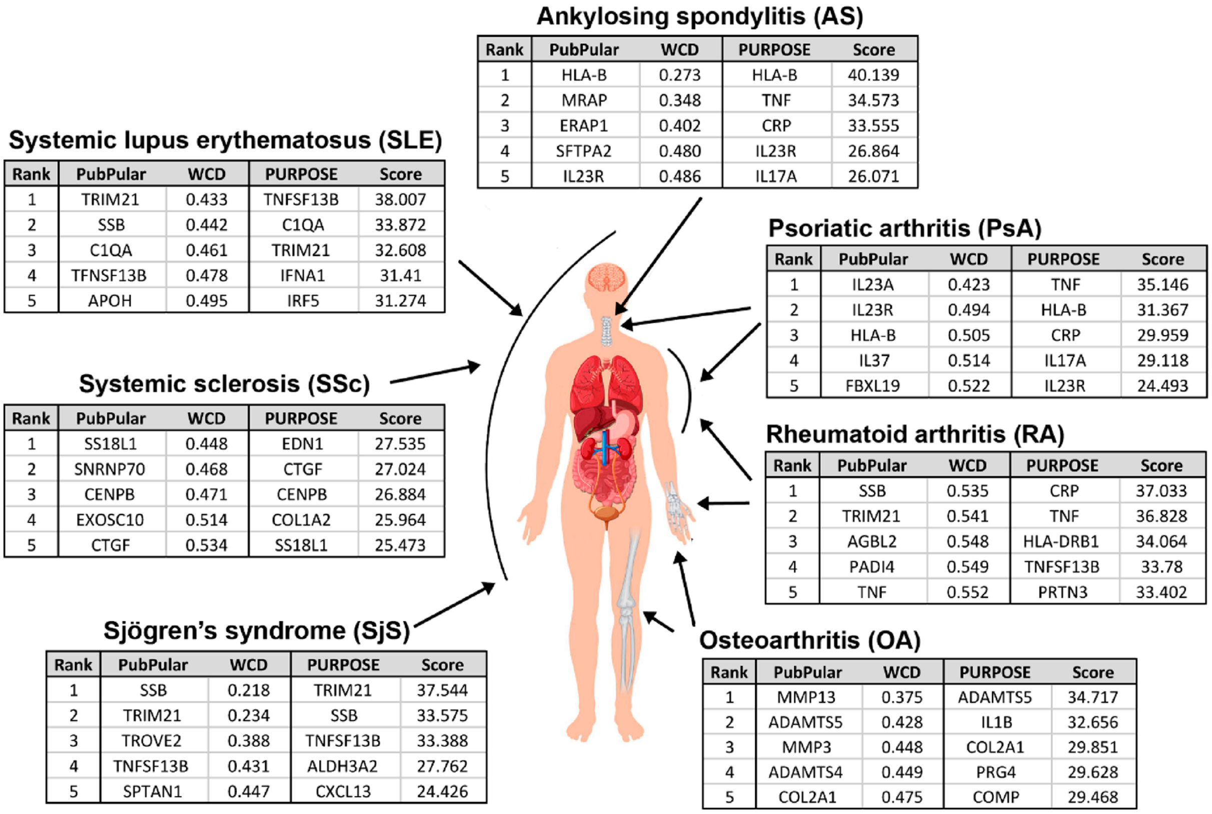
Figure 1. Popular proteins in rheumatic and autoimmune diseases. (A) Absolute number of proteins related to the field of RADs, according to the PURPOSE tool, and the processes in which they are involved. (B) Number of publications associated with each of the seven representative RADs included in this study, in both human (left) and mouse models (right). (C) Number of proteins included in the retrieved publications. OA, osteoarthritis; RA, rheumatoid arthritis; PsA, psoriatic arthritis; AS, ankylosing spondylitis; SLE, systemic lupus erythematosus; SSc, systemic sclerosis; and SjS, Sjögren's syndrome.

of the people who suffer them. In addition, many of these pathologies are increased in an aging population, so it is expected that their socioeconomic impact will further increase in the coming years.

Although the exact etiology of these diseases has not been fully established in many cases, it is known that many RMDs are autoimmune disorders (ADs) and, more specifically, autoimmune connective tissue disorders (ACTDs). Many of these cause substantial morbidity and mortality in patients, such as RA, SLE, and systemic sclerosis (scleroderma, SSc). The RAD initiative of the Human Proteome Project (RAD-HPP) was launched at the 2017 International HUPO meeting in Dublin, Ireland. The aim of the initiative is to tackle the unmet clinical needs in RADs, such as the development of improved diagnostics, the identification of novel drug targets, the establishment of targeted interventions, and improvement in clinical management using proteomics and its allied OMICs approaches. One of the immediate scientific goals of this initiative, under the frame of the biology-and-disease-centric strategy of the HPP (B/D-HPP),[4,5] is to assemble prioritized lists of proteins clinically relevant in RADs using the so-called "popular proteins" strategy and text mining software.[6,7] Remarkably, none of the tissues most affected in arthritis, such as articular cartilage, synovial tissue, or bone, are currently included in the Human Protein Atlas, even though such efforts have been initiated.[8] Furthermore, available proteomic data are limited to the human bone and pig/horse synovial fluid from the PeptideAtlas repository.[9] In addition to this lack of information in the field of proteomics, these tissues are also not represented in gene expression databases such as the Genotype-Tissue Expression (GTEx) portal.[10]

Literature records in the PubMed database currently exceed 29 million as of November 2018, of which 17.64 million are associated with proteins. Recently developed tools, such as

DOI: 10.1021/acs.jproteome.9b00360
J. Proteome Res. 2019, 18, 4231−4239



Ankylosing spondylitis (AS)

Rank	PubPular	WCD	PURPOSE	Score
1	HLA-B	0.273	HLA-B	40.139
2	MRAP	0.348	TNF	34.573
3	ERAP1	0.402	CRP	33.555
4	SFTPA2	0.480	IL23R	26.864
5	IL23R	0.486	IL17A	26.071

Systemic lupus erythematosus (SLE)

Rank	PubPular	WCD	PURPOSE	Score
1	TRIM21	0.433	TNFSF13B	38.007
2	SSB	0.442	C1QA	33.872
3	C1QA	0.461	TRIM21	32.608
4	TFNSF13B	0.478	IFNA1	31.41
5	APOH	0.495	IRF5	31.274

Psoriatic arthritis (PsA)

Rank	PubPular	WCD	PURPOSE	Score
1	IL23A	0.423	TNF	35.146
2	IL23R	0.494	HLA-B	31.367
3	HLA-B	0.505	CRP	29.959
4	IL37	0.514	IL17A	29.118
5	FBXL19	0.522	IL23R	24.493

Systemic sclerosis (SSc)

Rank	PubPular	WCD	PURPOSE	Score
1	SS18L1	0.448	EDN1	27.535
2	SNRNP70	0.468	CTGF	27.024
3	CENPB	0.471	CENPB	26.884
4	EXOSC10	0.514	COL1A2	25.964
5	CTGF	0.534	SS18L1	25.473

Rheumatoid arthritis (RA)

Rank	PubPular	WCD	PURPOSE	Score
1	SSB	0.535	CRP	37.033
2	TRIM21	0.541	TNF	36.828
3	AGBL2	0.548	HLA-DRB1	34.064
4	PADI4	0.549	TNFSF13B	33.78
5	TNF	0.552	PRTN3	33.402

Sjögren's syndrome (SjS)

Rank	PubPular	WCD	PURPOSE	Score
1	SSB	0.218	TRIM21	37.544
2	TRIM21	0.234	SSB	33.575
3	TROVE2	0.388	TNFSF13B	33.388
4	TNFSF13B	0.431	ALDH3A2	27.762
5	SPTAN1	0.447	CXCL13	24.426

Osteoarthritis (OA)

Rank	PubPular	WCD	PURPOSE	Score
1	MMP13	0.375	ADAMTS5	34.717
2	ADAMTS5	0.428	IL1B	32.656
3	MMP3	0.448	COL2A1	29.851
4	ADAMTS4	0.449	PRG4	29.628
5	COL2A1	0.475	COMP	29.468

Figure 2. Top five most popular proteins in each of the seven representative RADs included in this study, according to the PubPular v3.1 and PURPOSE tools.

PubPular[6] and the Protein Universal Reference Publication-Originated Search Engine (PURPOSE),[11] allow the systematic identification and prioritization of proteins related to a topic of interest. In this study, we apply these publicly available literature-mining platforms to focus on the field of RADs. The identification of proteins most strongly associated with RADs ("popular" or "high-priority" proteins in this topic) not only will enable the systematic prioritization of candidates to perform targeted proteomics studies but also would allow a greater insight into the key pathogenic processes of these disorders. Furthermore, it would enable the interrogation of those proteins, which may be shared between RADs and other conditions (e.g., cardiovascular or infectious).

2. MATERIALS AND METHODS

2.1. Literature Mining Software

To identify the most commonly associated proteins in the field of RADs, we employed two data mining software programs/algorithms: PubPular and PURPOSE. The PubPular tool determines the relevance of a protein in a topic of interest by the calculation of the normalized copublication distance (NCD, PubPular v2.3)[6] and, in a more recent version, the weighted copublication distance (WCD, PubPular v3.1),[12] whereas the PURPOSE method prioritizes the proteins according to their protein publication score (PURPOSE score).[11] Both algorithms take into account the number of publications related to the protein and topic of interest.

2.2. Keywords

The strategy was addressed in two steps: first, searching for high-priority proteins in specific representative RADs; then, evaluating shared hits between different RADs. To interrogate disease-specific publications, we used the following terms: "autoimmune disease", "rheumatic disease", "autoimmune connective tissue disease", "osteoarthritis", "rheumatoid arthritis", "spondyloarthropathy", "ankylosing spondylitis", "psoriatic arthritis", "systemic lupus erythematosus", "systemic sclerosis", and "Sjögren's syndrome". Whereas only human proteins were ranked with PubPular, the searches in the PURPOSE tool were performed for both human (*Homo sapiens*) and mouse (*Mus musculus*). The lists of popular proteins associated with RADs have been uploaded to PeptideAtlas and can be accessed via http://www.peptideatlas.org/PASS/PASS01449.

To evaluate the overlap with other HPP initiatives, we queried both in PubPular and PURPOSE using the terminology included in the "HPP Targeted Area" of the latter engine: "rheumatic" for the RAD initiative, "cardiovascular" for the cardiovascular initiative, "infectious OR infection" for the infectious diseases initiative, "immune OR immune system" for the immune peptidome initiative, and "muscle OR bone OR musculoskeletal" for the musculoskeletal initiative.

2.3. Data Analysis

Proteins were found for each topic, as previously mentioned, and identified using their UniProt accessions. For postanalyses, Gene Ontology terms associated with the identified proteins using the QuickGO tool were retrieved from the European Bioinformatics Institute (https://www.ebi.ac.uk/QuickGO/). The STRING v 10.5 tool (https://string-db.org/) was employed to visualize functional protein association networks and further GO analyses. Significant enrichment of annotations in a protein list over the background was calculated with the

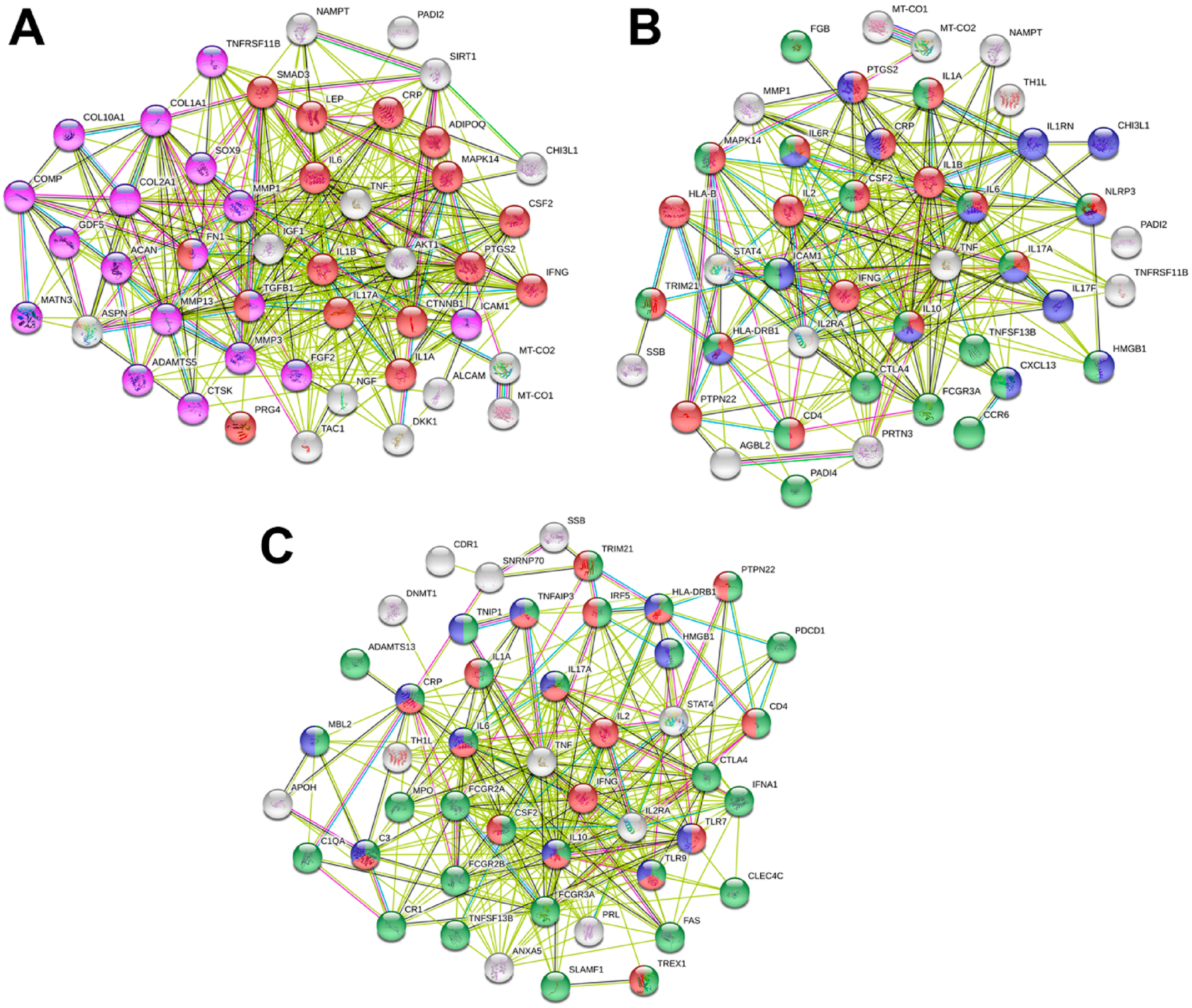


Figure 3. Functional networks built with popular proteins in three representative RADs. (A) Osteoarthritis, (B) rheumatoid arthritis, and (C) systemic lupus erythematosus. To facilitate the visualization, the figure shows only representative results obtained with the top 50 ranked proteins in each disease using the PURPOSE engine. Colors refer to the biological processes in which these proteins participate: pink, ECM organization; red, cytokine production; blue, inflammatory response; and green, immune system process. Functional networks obtained for the other four diseases analyzed in this study are provided in Supplementary Figure S1.

adjustment of the false discovery rate using the Benjamini−Hochberg method. The data were visualized using different Venn diagram tools: Venn diagram (http://bioinformatics.psb.ugent.be/webtools/Venn/), BioVenn (http://www.biovenn.nl/), and InteractiVenn (http://www.interactivenn.net/index2.html).

3. RESULTS

The results presented here are based on an analysis of the protein-associated data as of January 2019 from around 17.6 million publications and using the keywords defined in the Materials and Methods section.

3.1. Proteins Associated with Rheumatic and Autoimmune Disorders

439 629 papers were identified by literature mining up until January 2019 that collectively describe 4328 proteins in the context of ADs and 212 524 papers describing 2810 proteins in the context of rheumatic diseases (RDs). Although the searches were performed in all cases with the two literature mining engines, Figure 1 illustrates the results from PURPOSE because only this tool allowed the analysis on species different from human. As shown in Figure 1A, AD and RD shared almost 2500 proteins, which mainly participate in immune and inflammatory processes. A reduced number of proteins (337, 12%) were specifically related to RD in the literature, with no overlap with autoimmunity. Proteins included in this latter group were primarily related to the biology of connective and articular tissues or associated with OA, which is the most prevalent RD and is not an AD. Completely embedded between RD and AD is the term "ACTDs", which included 1248 proteins that participate in immune and inflammatory processes. The segregation of the publications and identified proteins into representative RADs is illustrated in Figure 1B,C. The highest numbers of proteins were identified for RA, SLE, and OA. The PubPular tool identified a number of proteins

associated with RADs, approximately two times higher compared with PURPOSE in all searches but retaining the same proportions between diseases (where RA is the pathology with the highest number of associated proteins and PsA is the pathology with the lowest number). Apart from human, the most studied animal model has been the mouse, with up to 796 murine proteins identified that are associated with RA. For this disease, mouse models include collagen-induced arthritis (CIA, an active immunization strategy) and also antibody-induced arthritis models (passive immunization strategies), such as collagen antibody-induced arthritis (CAIA) and K/BxN antibody transfer arthritis.[13] As shown in Figure 1B,C, the proportion between human and murine protein identifications in RADs (around one-third in mouse compared with what is reported in humans) is maintained for all of the representative diseases except psoriatic arthritis (PsA) and ankylosing spondylitis (AS), in which the amount of murine proteins is significantly lower. Animal models for these two diseases are less commonly used and generally restricted to HLA-27 transgenic rats or SKG mice (a monogenic model of autoimmune arthritis).[14]

We next used the PubPular v3 and PURPOSE tools to rank the proteins specifically related to each of these representative RADs as "popular" or "high-priority" proteins. Figure 2 shows the five most popular proteins for each of these disorders obtained with the two different engines, which gives an initial indication of the primary pathways and functions related to each pathological process. The top 100 proteins in each disease are listed in Supplementary Tables S1 (PubPular) and S2 (PURPOSE). These lists have been uploaded to PeptideAtlas, where they can be accessed with the data set identifier PASS01449.

3.2. Popular Proteins in Osteoarthritis

OA is the most prevalent RD, affecting around 10% of men and 18% of women above 60 years of age. According to WHO data, OA is the sixth leading cause of years lived with disability worldwide.[15] Furthermore, the number of people affected with symptomatic OA is increasing due to the aging of the population and the obesity epidemic.[16]

OA is a disorder involving movable joints characterized by cell stress and extracellular matrix (ECM) degradation initiated by micro- and macro-injury that activates maladaptive repair responses including pro-inflammatory pathways of innate immunity.[17] Although OA involves many different joint components, the most characteristically affected tissue is articular cartilage.[18] Consequently, the top 100 popular proteins in this disease (Supplementary Tables S1 and S2) include several proteins related to cartilage ECM, such as collagens (including types I, II, IX, X, and XI), aggrecan, and other ECM-related proteins (COMP, MATN3, PRG4, CILP, and UCMA). Within this, catabolic epitopes of type II and type I collagen (the most abundant proteins in the ECM of articular cartilage and bone, respectively) are among the most studied molecules as biomarkers for OA.[19,20] As shown in Figure 2, the top five ranked proteins in OA using PubPular include type II collagen, two matrix metalloproteinases (MMP-13 and -3), and two aggrecanases (ADAMTS-4 and -5). With PURPOSE, interleukin-1β (IL1B) and two proteins that have been proposed as biomarkers of OA, COMP[21] and PRG4 (also known as lubricin or superficial zone protein (SZP)), also appeared.[22]

Among the other groups of proteins most represented in this top-100 list are bone morphogenetic proteins, differentiation factors, and proteins related to inflammatory processes. Gene ontology analysis showed an enrichment of processes related to ECM organization and disassembly, collagen catabolism, ossification, and the development of the skeletal system, cartilage, or connective tissues (Supplementary Table S3). This can also be visualized through the study of functional protein association networks, which are shown in Figure 3. Proteins involved in ECM organization processes are highly represented in OA (Figure 3A, colored in pink), in contrast with those seen in two other very diverse RADs, RA (which has a mainly inflammatory phenotype, colored in blue in Figure 3B) and SLE (with a key autoimmune component, green in Figure 3C). The functional networks obtained for the remaining four representative RADs analyzed in this work are illustrated in Supplementary Figure S1.

3.3. Popular Proteins in RA

2361 human proteins were identified as associated with RA. This chronic AD affects between 0.5 and 1% of adults in the developed world, with between 5 and 50 new cases per 100 000 people per year.[23] Although it has a primary effect on joints, RA can also cause inflammation in organs such as the eyes, lungs, or kidneys, and it is highly disabling.[24]

The lists of the top 100 popular proteins in RA by PubPular and PURPOSE are enumerated in Supplementary Tables S1 and S2. Among them, the top five proteins share only tumor necrosis factor alpha (TNF) (Figure 2). Remarkably, the PubPular tool situates Ro (SSB) and La (TRIM21/SSA) autoantigens, which are mostly associated with SjS and SLE, as the top ranked priority proteins in RA. They are followed by AGBL2, a cytosolic carboxypeptidase[25] whose link with RA seems to be currently unknown, and PADI4, a peptidyl arginine deiminase that catalyzes the posttranslational conversion of the amino acid arginine to citrulline in the context of flanking linear amino acid sequences.[26−28] This posttranslational process is as an essential component of inflammation in a variety of diseases because of its role in inducing anticitrullinated protein/peptide antibodies (ACPAs), a class of autoantibodies that are of diagnostic, predictive, and prognostic value for RA.[29,30] In contrast with the more unexpected proteins obtained with PubPular (such as SSB, TRIM21, and AGBL2), the PURPOSE tool results in five proteins traditionally related to this disease, related to inflammation (CRP, TNF, TNFSF13B), immunity (HLA), and neutrophil function (PRTN3).

3.4. Popular Proteins in Spondyloarthropathies: Psoriatic Arthritis and Ankylosing Spondylitis

The term spondyloarthropathy (SpA) refers to any joint disease of the vertebral column, with lower back pain and stiffness being the most common clinical presentation. Often termed seronegative spondylarthropathies, this refers to the fact that they are negative for rheumatoid factor (RF) and ACPA, indicating a different etiopathogenesis to RA. Several RDs are included in this group, with AS and PsA being the two most prevalent SpAs.

The lists of the top 100 popular proteins in spondyloarthropathies identified by PubPular and PURPOSE are enumerated in Supplementary Tables S1 and S2. AS and PsA share some common proteins, independently of the literature mining tool that has been employed (Figure 2). These include an HLA antigen and the IL-23 receptor

(IL23R), which is clearly attributable to the increased incidence of the HLA-B27 allele in SpAs[31] and the known role of the IL23/IL17 axis in the pathogenesis of these diseases.[32] Interestingly, as happened in RA, some unusual hits appear with PubPular, such as FBXL19 (a component of the ubiquitin ligase complex) in PsA and SFTPA2 (a surfactant protein) in AS.

3.5. Popular Proteins in Systemic Autoimmune Connective Tissue Disorders

ACTDs are characterized as a group by the presence of an abnormal immune response that causes systemic damage to connective tissues. ACTDs include SLE, SSc (or scleroderma), Sjögren's syndrome (SjS), autoimmune myositides, overlap syndromes, and mixed connective tissue disease (MCTD).

We focused on three different representative ACTDs: SSc (primarily affecting skin and blood vessels), SLE (which may involve many organs and systems), and SjS (characterized by lymphocytic infiltrates in exocrine organs). The top 100 popular proteins found for each of these pathologies by PubPular and PURPOSE are enumerated in Supplementary Tables S1 and S2. Around one-third of these top 100 were common between the three diseases. Two expected candidates, TRIM21 (SSA or Ro protein) and SSB (Lupus La protein), were ranked at the top in both SLE and SjS (Figure 2), although SSB fell to the seventh position using PURPOSE (Supplementary Table S2). In SSc, the two literature mining software also included widely known proteins into their top five, such as CENPB (the primary target of the B cell anti-CENP response in SSc[33]) and CTGF (with a well-documented involvement in SSc fibrosis[34]). Nevertheless, apart from these anticipated hits, the search provided several proteins whose associations with ACTDs remain not fully understood, such as the beta-2-glycoprotein 1 (APOH) in SLE, the small nuclear ribonucleoprotein (SNRNP70) in SSc, or the chemokine 13 (CXCL13) in SjS. A targeted analysis of these particular proteins is needed to confirm any association identified by the present literature mining approach.

3.6. Overlap with Other HPP Initiatives

Apart from the debilitating mobility and pain, RAD patients are also at an increased risk of other pathologies. For example, several studies have shown an increased risk of serious infection in RA and other RADs, which could be explained by the pathobiologies of these diseases themselves, the impact of chronic comorbid conditions, or their immunosuppressive therapy.[35] Furthermore, it has been extensively reported that rheumatic patients are at an increased risk of cardiovascular events.[36] As illustrated in Supplementary Figure S2, whereas no significant overlap was found between these groups using the PubPular engine, the PURPOSE did show a number of proteins that can be associated with several HPP initiatives. The results from this analysis are listed in Supplementary Table S4: Three interleukins (IL-1A, -1B, and -6), the C-reactive protein, and MAPK14 were identified in the top 100 of these five initiatives when studied with PURPOSE. The largest overlap was found with the initiative of infectious diseases, where 17 proteins were shared among them, including several cytokines, for example, TNF, IFNA, IFNG, and IL2.

4. DISCUSSION

An exponential increase in the number of proteomic studies in recent years has led to a huge amount of data about proteins and their relationship to biology and disease. For this reason, there is currently paramount interest in managing and organizing this information as efficiently as possible to maximize the impact of proteomics data sets.

In the present study, we provide prioritized lists of proteins associated with rheumatic and autoimmune diseases. The fact that the tissues more characteristically affected in these pathologies, such as the components of the human joint, are missing in large-scale proteomic initiatives like the Human Protein Atlas or are scarcely represented in repositories such as PeptideAtlas turns the investigation of proteins associated with RADs into an essential effort both to improve the knowledge of poorly characterized proteins (uPE1)[37] and to facilitate the detection of "missing proteins" in the human proteome.[38] In this work, we establish a kick-off point for the systematic study of proteins associated with RADs, starting with the identification of those proteins that have been mostly referenced in the literature.

Interestingly, the overlap between the lists obtained from the two literature mining tools employed in this work is surprisingly low, ranging from just 23% for AS to 48% in the case of RA (Supplementary Figure S3 and Supplementary Table S5). Although this percentage has been increased using the WCD algorithm of the latest version of PubPular, it still remains a mean of 36% for all of the representative diseases examined. This result evidences the differences and putative complementarities of the two literature mining search engines that were utilized. As shown in Table 1 and Supplementary Table S6, the redundancy of protein occurrences among the representative RADs is much higher using the PURPOSE tool. In the latter case, eight proteins (including C-Reactive Protein, TNF, and four interleukins (IL1A, IL1B, IL6, and IL17A) appeared in all seven RADs (Table 1), whereas using the PubPular approach, the highest multiplicity was 5. One possible explanation for this result might be that the algorithm employed with PubPular[6] more strictly down-ranks proteins with high publication counts. This promotes query-specific proteins in the PubPular lists and diminishes the presence of proteins of general interest with large numbers of publications in multiple fields (such as CRP, TNF or ILs), which appear more often with PURPOSE.

Table 1. Redundant Proteins from the Top 100 High-Priority Proteins Observed in Four or More of the Seven Representative RADs That Have Been Studied in This Work

	7 RADs	6 RADs	5 RADs	4 RADs	
PURPOSE	IL17A TNF IL1A HLA-B IL1B IFNG CRP IL6	ACR ICAM1 HLA-DRB1 CD4 IL2RA IL10 FCGR3A CTLA4 IL2 KRT20	CSF2 FN1 PTPN22 FGB FANCB IFNA1 NELFCD PRTN3 CD24 TNFSF13B C1QA	MMP3 PTGS2 CSF1 TNFRSF11B CHI3L1 DKK1 COMP IL1RN IL23R HLA-DRB4	MPO TRIM21 IRF5 SSB STAT4 HLA-DQB1 PRL HLA-DPB1 CR2
PUBPULAR			IL17A TNFAIP3	MMP3 COMP AGBL3 AGBL2 BANK1 TRIM21 TNFSF13B IRF5 SSB	TROVE2 STAT4 HLA-DRB1 FAM167A BLK TNFSF13 TNIP1

Apart from these ubiquitous proteins that play well-known roles in either inflammatory or immune processes and could therefore be considered "positive controls" in the searches, the present literature mining approach also provides interesting lists of proteins that (1) had been described as related to the disease at only the genetic level, with publications reporting the association of specific polymorphisms with any of the pathologies, and (2) were first identified as antigens (through the detection of specific antibodies in the patient's sera) but whose putative role in the pathogenesis of the disease remains unclear. Although the literature mining approach followed in this work would probably determine the more commonly "popularity" of the identified proteins, these latter examples of molecules scarcely investigated at the protein level can be also considered high-priority proteins to perform further functional analysis to improve their characterization or elucidate their putative role as markers of disease.

Into the first group, we found ERAP1 to be associated with AS. This protein is an aminopeptidase specifically localized to the endoplasmic reticulum that trims peptides to their optimal size for binding to MHC proteins.[39] Interestingly, very recent studies have provided insights into ERAP1 polymorphisms, supporting the notion of using aminopeptidase inhibition to treat AS.[40,41] However, the multiplicity of ERAP1 variants and the distinct effects of their co-occurring polymorphisms require further efforts to elucidate their impact on the protein function and its relation to disease.

On the contrary, a representative example of the second group is the identification of EXOSC10 (exosome component 10) in SSc. This protein is also termed "Polymyositis-scleroderma overlap syndrome-associated autoantigen" or "Autoantigen PM/Scl-100". Anti-PM/Scl antibodies, first described as "anti-PM-1" in 1977, were found in patients with overlap syndrome of polymyositis (PM) and SSc. At a later date, the antigenic complex was identified as the human exosomes,[42] but only recent studies have described a putative malfunction of AS exosomes that may promote a profibrotic phenotype.[43] However, the participation of precise exosomal proteins in the molecular mechanisms underlying these processes should be further investigated as a source of novel specific biomarkers or drug targets.[44]

Altogether, the strategy based on literature mining tools that has been employed herein presents paramount advantages, such as the ease of use, enabling a straightforward prioritization and the collection of a large amount of data related to the topic of interest. However, this approach also has limitations provoked by the multiplicity of terms and abbreviations of proteins, which can lead to errors in the identifications. Because these mistaken identities were scarcely found in the searches, this supports the validity of the literature mining approach as a first step to find proteins related to disease but also underlines the need to review the results prior to the further analysis of specific proteins. A singular example of this problem is the protein SS18L1, which was identified by the two mining tools at the top of the list in SSc. This protein is a calcium-responsive trans-activator required for dendritic growth and branching in cortical neurons and thus from literature cannot be related to the disease. However, in this case, the engines were confounded by the alternative abbreviation of this protein, CREST (from calcium-responsive trans-activator), as the so-called CREST syndrome is a type of limited SSc whose acronym refers to its five main features: calcinosis, Raynaud's phenomenon, esophageal dysmotility, sclerodactyly, and telangiectasia. Finally, another essential limitation of the present literature mining approach relies on its bias toward proteins that have been well studied essentially because good assays exist for them, obviating other very important but less explored ones. In this sense, there are many proteins that should be of high priority despite the fact that they are unpopular.[45] Caution should be then taken into account when interpreting the results, considering the so-called "circular arguments", which essentially state that prior knowledge about function is biased toward well-studied genes or proteins.[46] Many predictions would thus be generic, so the most likely candidate proteins tend to be genes/proteins that have numerous other functions.[47] Altogether, the need for further targeted proteomic methodologies and workflows to enable the analysis of those less characterized proteins still remains a key challenge for the HPP initiatives.

In conclusion, the present work fulfils the first scientific goal of the RAD-HPP initiative by providing a map of the prioritized lists of proteins, which constitutes an initial step to perform further functional analyses on specific groups of proteins and pathologies. The coordinated effort of the RAD-HPP groups will be the key to further advancing in the characterization of the proteome associated with these pathologies and ultimately will aid in improving the management and quality of life of patients with RADs.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jproteome.9b00360.

Supplementary Figure S1. Functional networks built with the top 50 popular proteins in PsA, AS, SSc, and SjS. Supplementary Figure S2. Overlap of the top 100 popular proteins associated with RAD and other HPP initiatives using the two literature mining tools. Supplementary Figure S3. Percentages of common proteins identified in the top 100 with the two literature mining tools employed in this work (PDF)

Supplementary Table S1. Top 100 proteins in representative RADs using the PubPular tool (XLSX)

Supplementary Table S2. Top 100 proteins in representative RADs using the PURPOSE tool (XLSX)

Supplementary Table S3. Functional biological process-enrichment analysis of the top 50 most popular proteins in RADs. These data are illustrated in Figure 3 and Supplementary Figure S1 (XLSX)

Supplementary Table S4. Overlap of RAD-associated proteins with the top 100 popular proteins in other HPP initiatives. Data are illustrated in Supplementary Figure S2 (XLSX)

Supplementary Table S5. Overlap between the top 100 proteins that were identified as being associated with representative RADs using the two literature mining engines employed in this work (XLSX)

Supplementary Table S6. Overlap of the top 100 popular proteins identified in each of the seven representative RADs that have been analyzed in this study (XLSX)

■ AUTHOR INFORMATION

Corresponding Authors

*E-mail: cristina.ruiz.romero@sergas.es. Tel: 34-981-176399. Fax: 34-981-176398 (C.R.-R.).

*E-mail: fblagar@sergas.es. Tel: 34-981-176399. Fax: 34-981-176398 (F.J.B.).

ORCID

Cristina Ruiz-Romero: 0000-0001-7649-9803
Maggie P. Y. Lam: 0000-0001-9488-8319
Jennifer E. Van Eyk: 0000-0001-9050-148X
Justyna Fert-Bober: 0000-0002-2824-5056
Francisco J. Blanco: 0000-0001-9821-7635

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The Proteomics Unit of GIR belongs to ProteoRed/PRB3-ISCIII, supported by grant PT17/0019/0014. This work has been funded by grants from Fondo Investigación Sanitaria-Spain (PI16/02124, PI17/00404, DTS17/00200, CIBER-BBN CB06/01/0040, and RETIC-RIER-RD16/0012/0002), a part of the National Plan for Scientific Program Development and Technological Innovation 2013-2016, funded by the ISCIII-General Subdirection of Assessment and Promotion of Research – European Regional Development Fund (FEDER) "A way of making Europe".

■ REFERENCES

(1) van der Heijde, D.; Daikh, D. I.; Betteridge, N.; Burmester, G. R.; Hassett, A. L.; Matteson, E. L.; van Vollenhoven, R.; Lakhanpal, S.; et al. Common language description of the term rheumatic and musculoskeletal diseases (RMDs) for use in communication with the lay public, healthcare providers and other stakeholders endorsed by the European League Against Rheumatism (EULAR) and the American College of Rheumatology (ACR). *Ann. Rheum. Dis.* **2018**, *77* (6), 829–832.

(2) Norton, S.; Koduri, G.; Nikiphorou, E.; Dixey, J.; Williams, P.; Young, A. A study of baseline prevalence and cumulative incidence of comorbidity and extra-articular manifestations in RA and their impact on outcome. *Rheumatology (Oxford, U. K.)* **2013**, *52* (1), 99–110.

(3) Hoy, D. G.; Smith, E.; Cross, M.; et al. The global burden of musculoskeletal conditions for 2010: an overview of methods. *Ann. Rheum. Dis.* **2014**, *73* (6), 982–989.

(4) Aebersold, R.; Bader, G. D.; Edwards, A. M.; et al. The biology/disease-driven human proteome project (B/D-HPP): enabling protein research for the life sciences community. *J. Proteome Res.* **2013**, *12* (1), 23–27.

(5) Van Eyk, J. E.; Corrales, F. J.; Aebersold, R.; et al. Highlights of the Biology and Disease-driven Human Proteome Project, 2015–2016. *J. Proteome Res.* **2016**, *15* (11), 3979–3987.

(6) Lam, M. P. Y.; Venkatraman, V.; Xing, Y.; et al. Data-Driven Approach To Determine Popular Proteins for Targeted Proteomics Translation of Six Organ Systems. *J. Proteome Res.* **2016**, *15* (11), 4126–4134.

(7) Omenn, G. S. Advances of the HUPO Human Proteome Project with broad applications for life sciences research. *Expert Rev. Proteomics* **2017**, *14* (2), 109–111.

(8) Uhlén, M.; Fagerberg, L.; Hallström, B. M.; et al. Proteomics. Tissue-based map of the human proteome. *Science* **2015**, *347* (6220), 1260419.

(9) Kusebauch, U.; Deutsch, E. W.; Campbell, D. S.; Sun, Z.; Farrah, T.; Moritz, R. L. Using PeptideAtlas, SRMAtlas, and PASSEL: Comprehensive Resources for Discovery and Targeted Proteomics. *CurrProtocBioinforma* **2014**, *46*, 13.25.1–13.25.28.

(10) Carithers, L. J.; Moore, H. M. The Genotype-Tissue Expression (GTEx) Project. *Biopreserv. Biobanking* **2015**, *13*, 307–308.

(11) Yu, K.-H.; Lee, T-LM; Wang, C.-S.; et al. Systematic Protein Prioritization for Targeted Proteomics Studies through Literature Mining. *J. Proteome Res.* **2018**, *17* (4), 1383–1396.

(12) Lau, E.; Venkatraman, V.; Thomas, C. T.; Wu, J. C.; Van Eyk, J. E.; Lam, M. P. Y. Identifying High-Priority Proteins Across the Human Diseasome Using Semantic Similarity. *J. Proteome Res.* **2018**, *17* (12), 4267–4278.

(13) Caplazi, P.; Baca, M.; Barck, K.; et al. Mouse Models of Rheumatoid Arthritis. *Vet. Pathol.* **2015**, *52* (5), 819–826.

(14) Breban, M.; Araujo, L. M.; Chiocchia, G. Animal models of spondyloarthritis: do they faithfully mirror human disease? *Arthritis Rheumatol.* **2014**, *66* (7), 1689–1692.

(15) Cross, M.; Smith, E.; Hoy, D.; et al. The global burden of hip and knee osteoarthritis: estimates from the global burden of disease 2010 study. *Ann. Rheum. Dis.* **2014**, *73* (7), 1323–1330.

(16) Neogi, T.; Zhang, Y. Epidemiology of Osteoarthritis. *Rheum Dis Clin North Am.* **2013**, *39* (1), 1–19.

(17) Kraus, V. B.; Blanco, F. J.; Englund, M.; Karsdal, M. A.; Lohmander, L. S. Call for standardized definitions of osteoarthritis and risk stratification for clinical trials and clinical use. *OsteoarthrCartil* **2015**, *23* (8), 1233–1241.

(18) Luo, Y.; Sinkeviciute, D.; He, Y.; et al. The minor collagens in articular cartilage. *Protein Cell* **2017**, *8* (8), 560–572.

(19) Lotz, M.; Martel-Pelletier, J.; Christiansen, C.; et al. Value of biomarkers in osteoarthritis: current status and perspectives. *Ann. Rheum. Dis.* **2013**, *72* (11), 1756–1763.

(20) Watt, F. E. Osteoarthritis biomarkers: year in review. *OsteoarthrCartil* **2018**, *26* (3), 312–318.

(21) Verma, P.; Dalal, K. Serum cartilage oligomeric matrix protein (COMP) in knee osteoarthritis: a novel diagnostic and prognostic biomarker. *J. Orthop. Res.* **2013**, *31* (7), 999–1006.

(22) Ogawa, H.; Matsumoto, K.; Terabayashi, N.; Kawashima, K.; Takeuchi, K.; Akiyama, H. Association of lubricin concentration in synovial fluid and clinical status of osteoarthritic knee. *Mod. Rheumatol.* **2017**, *27* (3), 489–492.

(23) Scott, D. L.; Wolfe, F.; Huizinga, T. W. J. Rheumatoid arthritis. *Lancet* **2010**, *376* (9746), 1094–1108.

(24) Cross, M.; Smith, E.; Hoy, D.; et al. The global burden of rheumatoid arthritis: estimates from the global burden of disease 2010 study. *Ann. Rheum. Dis.* **2014**, *73* (7), 1316–1322.

(25) Sahab, Z. J.; Hall, M. D.; Me Sung, Y.; et al. Tumor suppressor RARRES1 interacts with cytoplasmic carboxypeptidase AGBL2 to regulate the α-tubulin tyrosination cycle. *Cancer Res.* **2011**, *71* (4), 1219–1228.

(26) van Boekel, M. A. M.; van Venrooij, W. J. Modifications of arginines and their role in autoimmunity. *Autoimmun. Rev.* **2003**, *2* (2), 57–62.

(27) Wang, S.; Wang, Y. Peptidylarginine deiminases in citrullination, gene regulation, health and pathogenesis. *Biochim. Biophys. Acta, Gene Regul. Mech.* **2013**, *1829* (10), 1126–1135.

(28) Suzuki, A.; Yamada, R.; Chang, X.; et al. Functional haplotypes of PADI4, encoding citrullinating enzyme peptidylarginine deiminase 4, are associated with rheumatoid arthritis. *Nat. Genet.* **2003**, *34* (4), 395–402.

(29) van Venrooij, W. J.; Pruijn, G. J. M. How citrullination invaded rheumatoid arthritis research. *Arthritis Res. Ther* **2014**, *16* (1), 103.

(30) Valesini, G.; Gerardi, M. C.; Iannuccelli, C.; Pacucci, V. A.; Pendolino, M.; Shoenfeld, Y. Citrullination and autoimmunity. *Autoimmun. Rev.* **2015**, *14* (6), 490–497.

(31) Colbert, R. A.; Tran, T. M.; Layh-Schmitt, G. HLA-B27 misfolding and ankylosing spondylitis. *Mol. Immunol.* **2014**, *57* (1), 44–51.

(32) Smith, J. A.; Colbert, R. A. Review: The Interleukin-23/Interleukin-17 Axis in Spondyloarthritis Pathogenesis: Th17 and Beyond. *Arthritis Rheumatol.* **2014**, *66* (2), 231–241.

(33) Mehra, S.; Walker, J.; Patterson, K.; Fritzler, M. J. Autoantibodies in systemic sclerosis. *Autoimmun. Rev.* **2013**, *12* (3), 340–354.

(34) Rice, L. M.; Ziemek, J.; Stratton, E. A.; et al. A longitudinal biomarker for the extent of skin disease in patients with diffuse cutaneous systemic sclerosis. *Arthritis Rheumatol.* **2015**, *67* (11), 3004–3015.

(35) Listing, J.; Gerhold, K.; Zink, A. The risk of infections associated with rheumatoid arthritis, with its comorbidity and treatment. *Rheumatology (Oxford, U. K.)* **2013**, *52* (1), 53−61.

(36) Baillet, A.; Gossec, L.; Carmona, L.; et al. Points to consider for reporting, screening for and preventing selected comorbidities in chronic inflammatory rheumatic diseases in daily practice: a EULAR initiative. *Ann. Rheum. Dis.* **2016**, *75* (6), 965−73.

(37) Paik, Y.-K.; Lane, L.; Kawamura, T.; et al. Launching the C-HPP neXt-CP50 Pilot Project for Functional Characterization of Identified Proteins with No Known Function. *J. Proteome Res.* **2018**, *17* (12), 4042−4050.

(38) Baker, M. S.; Ahn, S. B.; Mohamedali, A.; et al. Accelerating the search for the missing proteins in the human proteome. *Nat. Commun.* **2017**, *8* (1), 14271.

(39) Martín-Esteban, A.; Sanz-Bravo, A.; Guasp, P.; Barnea, E.; Admon, A.; López de Castro, J. A. Separate effects of the ankylosing spondylitis associated ERAP1 and ERAP2 aminopeptidases determine the influence of their combined phenotype on the HLA-B*27 peptidome. *J. Autoimmun.* **2017**, *79*, 28−38.

(40) Hanson, A. L.; Cuddihy, T.; Haynes, K.; et al. Genetic Variants in *ERAP1* and *ERAP2* Associated With Immune-Mediated Diseases Influence Protein Expression and the Isoform Profile. *Arthritis Rheumatol.* **2018**, *70* (2), 255−265.

(41) Sanz-Bravo, A.; Alvarez-Navarro, C.; Martín-Esteban, A.; Barnea, E.; Admon, A.; López de Castro, J. A. Ranking the Contribution of Ankylosing Spondylitis-associated Endoplasmic Reticulum Aminopeptidase 1 (ERAP1) Polymorphisms to Shaping the HLA-B*27 Peptidome. *Mol. Cell. Proteomics* **2018**, *17* (7), 1308−1323.

(42) Brouwer, R.; Vree Egberts, W. T. M.; Hengstman, G. J. D.; et al. Autoantibodies directed to novel components of the PM/Scl complex, the human exosome. *Arthritis Res.* **2002**, *4* (2), 134−138.

(43) Wermuth, P. J.; Piera-Velazquez, S.; Jimenez, S. A. Exosomes isolated from serum of systemic sclerosis patients display alterations in their content of profibrotic and antifibrotic microRNA and induce a profibrotic phenotype in cultured normal dermal fibroblasts. *Clin. Exp. Rheumatol.* **2017**, *35* (Suppl 106), 21−30.

(44) Zhu, T.; Wang, Y.; Jin, H.; Li, L. The role of exosome in autoimmune connective tissue disease. *Ann. Med.* **2019**, *51*, 101−108.

(45) Carter, A. J.; Kraemer, O.; Zwick, M.; Mueller-Fahrnow, A.; Arrowsmith, C. H.; Edwards, A. M. Target 2035: probing the human proteome. *Drug Discovery Today* **2019**.

(46) Pavlidis, P.; Gillis, J. Progress and challenges in the computational prediction of gene function using networks. *F1000Research* **2012**, *1*, 14.

(47) Gillis, J.; Ballouz, S.; Pavlidis, P. Bias tradeoffs in the creation and analysisof protein-protein interaction networks. *J. Proteomics* **2014**, *100*, 44−54.