Exploring High-Throughput Immunoassays for Biomarker Validation in Rheumatic Diseases in the Context of the Human Proteome Project

Manuel Fuentes, Cristina Ruiz-Romero, Sara Misiego, Pablo Juanes-Velasco, Alicia Landeira-Viñuela, Adrián Torres-Roda, Héctor Lorenzo-Gil, María González-González, Ángela P. Hernández, Lucía Lourido, Ronald Sjöberg, Elisa Pin, Javier de Las Rivas, José Manuel Sánchez-Santos, Peter Nilsson, and Francisco J. Blanco

ABSTRACT

Rheumatic diseases are high prevalence pathologies with different etiology and evolution and low sensitivity in clinical diagnosis. Therefore, it is necessary to develop an early diagnosis method which allows personalized treatment, depending on the specific pathology. The biology/disease initiative, at Human Proteome Project, is an integrative approach to identify relevant proteins in the human proteome associated with pathologies. A previously reported literature data mining analysis, which identified proteins related to osteoarthritis (OA), rheumatoid arthritis (RA), and psoriatic arthritis (PSA) was used to establish a systematic prioritization of potential biomarkers candidates for further evaluation by functional proteomics studies. The aim was to study the protein profile of serum samples from patients with rheumatic diseases such as OA, RA, and PSA. To achieve this goal, customized antibody microarrays (containing 151 antibodies targeting 121 specific proteins) were used to identify biomarkers related to early and specific diagnosis in a screening of 960 serum samples (nondepleted) (OA, n = 480; RA, n = 192; PSA, n = 288). This functional proteomics screening has allowed the determination of a panel (30 serum proteins) as potential biomarkers for these rheumatic diseases, displaying receiver operating characteristics curves with area under the curve values of 80–90%.

KEYWORDS:

Protein microarrays; Biomarkers; Osteoarthritis; Rheumatoid arthritis; Psoriatic arthritis; Human Proteome Project

1. INTRODUCTION

The high complex nature of osteoarthritis (OA) has hindered the development of tools for the precise evaluation and therapeutic management of this disease.[1–3] The "-omics" technologies have contributed to increase the knowledge on OA pathogenesis and have provided lists of molecules related to these pathologies.[4–6] Further validations studies are still needed to translate findings from "-omics" studies into clinical applications. Having this in mind, the combination of this molecular info from "-omics" studies with clinical info would be highly valuable to design and perform biomarkers validation based on immunoassays in a high-throughput format.[7]

Rheumatism is a nonspecific term used to describe any painful disorder affecting the loco-motor system including joints, muscles, connective tissues, and soft tissues around the joints and bones.[8] Rheumatologic disorders include multiple different pathologies: (i) aging-related pathologies such as OA; (ii) inflammatory diseases such as RA or PSA; (iii) several systemic autoimmune diseases such as systemic lupus erythematosus, Bechet's disease, and others.[9] Regarding social aspects, rheumatic and autoimmune diseases have a huge socio-economic impact given the high prevalence or morbidity of some of them. Furthermore, they present several currently unmet medical needs: their early diagnosis is often difficult, they have challenging heterogeneous clinical courses, and treatment choices are not yet personalized.

Among the missions of the HUPO Biology/Disease-driven Human Proteome Project [10,11] initiatives are to identify proteins in the human proteome associated with diseases and biological systems.[12] Currently, several B/D-HPP initiatives are active and making progress in translational biomedical research. One of them is the initiative focused on be Rheumatic and Articular Diseases Team of Biology and Disease driven branch of the Human Proteome Project (HPP) (RAD-HPP). It was launched in 2016 with the general objective, as the RAD-HPP initiative, of generation of novel knowledge about proteins involved in prognostic, therapeutic, and physiopathological issues taking place in rheumatic and autoimmune diseases.[13] This will allow improving the capacity for prognosis and early diagnosis, exploring potential therapeutic biomarkers, and generating basic knowledge on alternative disease pathogenesis not presently addressed by current therapies. As a strategic goal, this RAD-HPP initiative aims to improve the research

capacity of the participating groups, as well as to promote cooperative research in this area through the development of networking activities and joint research activities.

Currently, the RAD-HPP initiative has described an action plan focused on: (i) expand visibility and membership of RAD-HPP; (ii) definition of the proteome of human joint tissues (such as cartilage, synovial membrane, meniscus, and subchondral bone; (iii) definition of protein lists relevant in the physiology of the joint and systemic autoimmune processes; (iv) assemble prioritized clinically relevant proteins in RAD, to boost the development of quantitative proteomics assays for systemic analysis (targeted MS kits for MRM analysis, antibody-based approaches such as protein microarrays), also for further characterization of proteins as potential useful biomarkers for patient stratification, therapeutic management; (v) development of tools for PTMs characterization that drive autoimmunity, facilitating their use in understanding physiological and pathological roles, and finally to understand the link between autoimmunity and heightened cardiovascular risk.

Recently, several software tools have been developed, such as PubPular,[14] PURPOSE and metaPURPOSE,[15] GLAD4U,[16] and FACTA+,[17] which allow researchers to prioritize crucial proteins in various organ-systems and diseases from the biomedical literature. Information from these databases can be used to facilitate further targeted and/or functional proteomics studies. Here, these approaches have been employed to select a panel of protein candidates and a few candidates from our previous studies[18,19] in order to analyze the differential plasma protein profiles by high-throughput immunoassays in a cohort of 960 samples (OA, RA, PSA, controls). Hence, it is presented a serum screening of 960 nondepleted samples (480 OA, 288 PSA, 192 RA, 83 healthy controls) by a customized antibody array (containing 151 antibodies targeting 121 different proteins) with the major goal to explore the ability of high-content immunoassays to establish a panel of biomarker candidates useful for diagnosis and/or prognosis of articular diseases in the context of the RAD-HPP initiative.

2. MATERIALS AND METHODS

2.1. Materials

Acetone >98% (Panreac, Barcelona, Spain), 3-(2-aminoethylamino) propyl-methyl dimethoxysilane (MANAE) (Fluka, Steinheim, Germany), dimethyl sulfoxide (DMSO) (Merck Millipore, Billerica, USA), bis(sulfosuccinimidyl)-suberate (BS3), Nunc 384 clear flat well plates, SuperBlock Blocking Buffer, Microtiter Plate 96 Well/V Bottom, LifterslipTM coverslips (Thermo Scientific, Portsmouth, USA), Bovine Serum Albumin (BSA) >98%, NHS-PEG4-Biotin, Tween20 viscous liquid (polyoxyethylenesorbitan monolaurate), ampicillin sodium, Corning hybridization chambers (Sigma-Aldrich, St Louis, USA), peroxidase-AffiniPure Goat Anti-Human IgG, Fcγ Fragment Specific, Peroxidase-AffiniPure F(ab′)2 Fragment Goat Anti-Mouse IgG (H+L) (Jackson ImmunoResearch Laboratories, Baltimore, USA), TSA individual cyanine 3 Tyramide Reagent Pack (TSA) (PerkinElmer, Waltham, USA), slides Ground Edges 76 × 26 mm (LíneaLab, Badalona, Spain), Amersham Cy5-Streptavidin, 16-Array Chamber Covers (GE Healthcare, Buckinghamshire, UK), Goat Anti-Rabbit IgG (H+L) HRP Conjugate (BIO-RAD, California, USA), powdered concentrated skimmed milk (Central Lechera Asturiana, Granda-Siero, Spain).

Purified rabbit polyclonal antihuman antibodies against potential biomarkers (Supporting Table 1), previously described to be altered in osteoarthritic patients at different levels in cartilage (extracellular matrix (ECM), chondrocytes) and blood, were kindly provided by the Human Protein Atlas.

2.2. Patients

Nondepleted, peripheral blood, human serum samples corresponding to patients (age average: 68 years (range: 55–80 years) belonging to three different groups: osteoarthritis (OA), rheumatoid arthritis (RA), and healthy controls (C) were provided by the Biobank at the Institute for Biomedical Research of A Coruña (INIBIC). The samples were extracted and processed, after written informed consent was given by each donor, according to the guidelines of the local Ethics Committee (Comité Ético de Galicia, Galicia, Spain). The OA group consisted of 480 patients diagnosed with OA according to

the American College of Rheumatology (ACR) criteria.[20] The RA group comprised 192 patients diagnosed with RA following the ACR/European League Against Rheumatism (EULAR) criteria [20] The PSA group consisted of 288 diagnosed with PSA following the ACR/European League Against Rheumatism (EULAR). The control group included 83 samples from patients with no history of joint disease and nonradiographic OA.

2.3. Methods

2.3.1. Meta-Analysis of Popular Proteins. To define a panel of protein candidates as possible biomarkers in PubPular is a R/Shiny web interface for querying and visualizing popular proteins for custom search terms using Gene2Pubmed/Pub Tator data sources and semantic similarity metrics as described by Lau et al.[14] The query search was performed in the updated version which included precompiled protein lists from Disease Ontology and Human Phenotype Ontology disease terms as well as reverse protein-to-topic searches. PubPular was accessed via https://maggielab.org/tool_pubpular.

The PURPOSE database (Protein Universal Reference Publication-Originated Search Engine) tool prioritizes proteins by the strength and specificity of the associations between proteins and the query terms. For each query, the number of PubMed publications are retrieved in real-time as described by Yu et al.[15] PURPOSE is accessible through http://rebrand.ly/proteinpurpose.

In both databases, the search queries were: osteoarthritis, arthritis, rheumatoid arthritis, psoriatic arthritis, serum, plasma, ... Among the list of proteins selected from these databases were several proteins from our previous reported studies based on affinity proteomics.[19] In general, most of the protein candidates are involved in featured biological processes described in these pathologies, such as matrix regeneration, cellular differentiation, and also proteins which belong to an inflammation or a humoral immune response as it may be expected in these pathologies.

2.3.2. Design and Performance of Antibody Microarray. The glass slide surfaces were activated[21] by treatment with 2% (v/v) MANAE in acetone for 30 min with shaking at room temperature (RT). Slides were subsequently washed with acetone and Milli-Q water and dried with compressed filtered air. In this case, antihuman polyclonal rabbit antibodies (from Human Protein Atlas, Supplementary Table 1) were resuspended (1:1,

v/v) in a 47% (v/v) glycerol solution, according to the ArrayJet Printer Marathon v1.4 specifications. NHS-PEG4-biotin (0.39 mg/mL) and BSA (concentrations ranging from 0.6 mg/L to 3.66 mg/mL) were prepared as positive and negative controls, respectively. Slides printed using ArrayJet Printer Marathon v1.4 (ArrayJet, Roslin, UK) contained 12 identical subarrays, each one including all antibodies and controls to be analyzed. The spot diameter was set at 100 μm and the separation distance among spots at 200 μm (Supplementary Figure 1). A total of 6 serum samples were analyzed per array, in duplicate. Eventually, printed arrays were packed and stored, protected from light in a dry atmosphere at RT, until assayed.[22]

2.3.3. Evaluation of Array Performance. All the following steps were performed at RT. Antibody arrays were blocked with a SuperBlock-PBS solution for 1 h on a rocking platform. Then, they were washed (5 min, 3×) with PBS 1×. After that, the arrays were incubated with HRP-conjugated antirabbit secondary IgG (1:200 (v/v) in SuperBlock-PBS) for 1 h in a humidified chamber. Then, they were individually washed with (i) PBS (5 min, 3×) and (ii) distilled water (5 min, 1×). Subsequently, arrays were incubated with a 1:50 (v/v) TSA solution for 10 min in a humidified chamber. Arrays were then washed as described above and dried with filtered compressed air. Finally, they were scanned using the GenePix 4000B Scanner (Axon Instruments, Union City, USA) and the SensoSpot Fluorescence Scanner (Sensovation AG; Radolfzell, Germany), and analyzed.

2.3.4. Sera Biotinylation. According to the protocol described by Sierra-Sanchez et al.[19] adapted from Häggmark et al.,[23] proteins present in OA, RA, and C sera samples were biotinylated by incubation with 0.78 mg/mL NHS-PEG4-biotin for 2 h at 4°C. Biotinylation reactions were stopped by adding 0.5 M Tris–HCl (pH 8).

2.3.5. Detection of Protein Serum Profiles. Antibody arrays were blocked with Superblock for 1 h with mild stirring and, subsequently, washed with distilled water (5 min, 3×). Then, 40 μL of a 1:1000 (v/v) dilution of biotinylated serum was added to each well of the 16-array chamber. Chambers were covered and incubated O/N at 4 °C with slight shaking. After that, the arrays were individually washed with distilled water and

revealed using a 1:50 (v/v) dilution of Cy5-Streptavidin for 20 min at RT. Finally, arrays were washed, dried with compressed filtered air, and scanned.[18]

2.3.6. Image Acquisition and Signal Standardization. The TIFF images generated by array scanning were analyzed using GenePix Pro 4.0. software. Parameters were set to quantify light intensity values at Cy3 (λ = 532 nm) and Cy5 (λ = 635 nm) emission wavelengths, respectively. Signal intensity values were normalized following eq 1, where S_i^N is referred to the intra- and inter-array normalized signal, S_i to the raw signal in antibody containing spots, S_{bg} to the the background signal, and S' to the raw signal in companion buffers' spots. Background signal subtraction within each array was followed by fold change calculation with respect to a blank across array.[22]

$$S_i^N = \frac{S_i - S_{bg}}{\text{median}\{S' : S' - S_{bg} > 0\}} \quad (1)$$

As several serum samples were assayed per array (in duplicate), the background signal was estimated as the mean of DEPC water spot intensity values present in each pair of subarrays corresponding to the same serum. The blank signal was estimated as the median of intra-array normalized intensity values of positive control spots containing the companion buffers (i.e., glycerol + PBS + BS3), across all the arrays.[18,19,22,24] The median signal of the replicates per antibody was computed as an estimate of reactivity. Antibodies with normalized intensity values greater than 1 were considered as hits (signal due to antibody–antigen binding) and the corresponding samples were considered as positive for the antigen specifically bound by the screened antibody.[18,19,22,24] The resulting intensity distribution is plotted in Supplementary Figure 1 (where every line represents a sample). Positive spots are normalized and further filtered by comparison with negative controls (i.e., BSA, DEPC water, BS3, and combinations).[18,19,22,24] After standard scoring was applied, the distribution seemed homogeneous enough for comparison without further distorting the data (Supplementary Figure 1). The significant level was 0.05.

2.3.7. Data Analysis Workflow. For the data analysis, to remove the effect of outlier samples, the sera in which the number of hits detected was below 1% or above 99% were not considered for further analyses.[25–27] The data analysis pipeline is depicted in Supplementary Figure 3. According to analyzed samples (Table 1), the cohort frequency distribution is observed in Figure 1A. In all these samples, a total of 121 proteins were simultaneously evaluated by the customized protein array containing 151 different antibodies. Then, 302 variables were included in the analysis among the groups of patients. In order to detect outliers, histograms and boxplots were analyzed according to R software with the RSudio interface.

For comparison between independent groups, the number of required pair comparisons was determined:

$$C_{(n,r)} = \binom{n}{r} = \frac{n!}{r!\,(n-r)!} 0 \le r \le n \tag{2}$$

where C = number of comparisons, n = number of groups of patients, r = number of groups per comparison. So, for r = 2 groups the number of comparisons is $n(n-1)/2$. First, for each pair of comparisons, we check the homogeneity of variances s_i^2 with the Bartlett test.

$$U = \frac{1}{C}[(N-k)\ln(s^2) - \sum_{i=1}^{k}(n_i - 1)\ln s_i^2]$$
$$\text{where} C = 1 + \frac{1}{3(k-1)}\left(\sum_{i=1}^{k}\frac{1}{n_i - 1} - \frac{1}{N-k}\right) \tag{3}$$

For the pairwise comparisons with homogeneous variances, we studied the differences in intensity using the Student *t*-test.

$$t_{n_1+n_2-2} = \frac{\bar{x}_1 - \bar{x}_2}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}, \text{where} S_p^2 = \frac{(n_1-1)s_{c_1}^2 + (n_2-1)s_{c_2}^2}{(n_1-1)+(n_2-1)} \tag{4}$$

For pairwise comparisons with heterogeneous variances, we used the Welch *t*-test.

$$t_g = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_{c_1}^2}{n_1} + \frac{s_{c_2}^2}{n_2}}}, \text{ where degrees of freedom is } g = \frac{(\frac{s_{c_1}^2}{n_1} + \frac{s_{c_2}^2}{n_2})^2}{\frac{(\frac{s_{c_1}^2}{n_1})^2}{n_1 + 1} + \frac{(\frac{s_{c_2}^2}{n_2})^2}{n_2 + 1}} - 2 \quad (5)$$

where $\bar{x}_i$ s_{ci}^2, and n_i are respectively the sample mean, the sample variance and the size of each group (i = 1, 2). The workflow for all this process is described in Supplementary Figure 2.

The differential protein profiles are observed in all side-by-side comparisons, these protein profiles were analyzed from the functional point of view in order to identify any featured biological role in these pathologies of interest. The Functional Enrichment Analysis (FEA) was performed by DAVID (David Bioinformatics Resources vs 6.8)[28] and Cytoscape.[29]

Regarding clustering analysis, it is performed by Ward Method with 4 groups corresponding with the pathologies individually (OA, RA, PSA) vs controls, and all pathologies (all of them as one single group) vs controls. In all the data analysis, several biostatistical approaches (Ward linkage, camberra distance, and silhouette average used for *k*-means cluster; logistic regression, multiclass random forest and receiver operating characteristics curve (ROC) all used for classification) were performed with R software using the RStudio interface.

3. RESULTS AND DISCUSSION

3.1. Evaluation of Array Performance

The array content was based on a previous search on the databases PubPular and Protein PURPOSE[14] (supported by HUPO), which has been reported and described by Ruiz-Romero et al.[13] In this study, these meta-data have been explored to select a panel of protein candidates for evaluating the presence in plasma by affinity proteomics. At first glance, the revised literature, by both databases, displays a homogeneous distribution (see

Supplementary Figure 3) across all the pathologies of interest. Thus, the selection of proteins is unbiased to all these rheumatic pathologies, and selected proteins have not been enriched in a particular functional group and/or pathologies. In this panel of protein candidates (Supplementary Table 1), TNF is the protein with a high number of citations and previous references in the reviewed literature. This fact might be expected as it is an inflammatory cytokine with systemic effects which is related to many physiopathologies.[30,31] Mostly of the protein candidates belong to the innate and specific immune response such as cytokines (IL6, IL10, ...), complement system (C3, C4, ...), histocompability complex (HLA-A, HLA-B, HLA-C, ...), immunophenotyping (CD14, CD68, ...), adhesion molecules (ICAM, CCL12, ...), and to bone matrix and metabolism such as matrix metallopeptidases or collagen. Moreover, several identified proteins in our previous reported studies have been also included in this panel for the further evaluation in serum by high-throughput immunoaffinity assays.

The 960 nondepleted serum samples were screened by using 120 arrays (8 serum samples per array) targeting 121 proteins by 151 different antibodies in each subarray. The cohort description displays 833 pathological serum samples and 127 serum samples from healthy controls. The absolute frequency is higher for the OA group (n = 433), followed by psoriatic arthritis (PSA) (n = 258) and RA (n = 192) (Figure 1A). Despite the high heterogeneity between pathological groups, and after signal normalization (as described in the Materials and Methods section), the data set displays a global normal distribution (Figure 1B), which is required for deciphering differential protein profiles across the analyzed cohort. Then, all targeted proteins (n = 121) were considered in all analyzed samples (n = 966) for the further analysis and could be potential candidates for panel of biomarkers. Once protein arrays were normalized and standardized (Figure 1C), overall differences were observed on the analyzed proteins across the pathological groups (Figure 1D).

3.2. Identification of Differential Protein Profiles

In order to establish differential protein profiles among the 4 groups of patients screened in this study, 6 possible comparisons were performed as described in Supplementary Figure 3, starting between healthy controls vs pathological groups and followed by evaluation of differences within pathological groups.

3.2.1. Pathological Groups vs Controls. *3.2.1.1. OA vs Controls***.** In Figure 2 is depicted a global view of protein profiles, from which is observed 16 different protein profiles according to 16 differential protein profiles, being 10 of them at a high-relative abundance level (Figure 2A) and 6 of them at a low-relative abundance level (Figure 2D). The top observed proteins are in the high-relative abundance level in OA, as might be expected; in addition, there is a match of several of these highlighted proteins and the ones identified by PURPOSE and PubPular databases (such as TNF, APOM, C1Q, COMP, among others). Mostly of these proteins are related with extracellular matrix organization (i.e., APOM, COMP, FBN1, ...), innate immune response (i.e., ANG, C1Q, FBNL1, ...). However, low relative abundance levels are observed for proteins related to platelet activation/signaling (i.e., APBB1IP, CLEC3, ...), complement cascade (i.e., CFH, FC1, ...), and structural and metabolic pathways (i.e., CSPG4, FSTL1, ...).

*3.2.1.2. PSA vs Controls***.** Here, 258 PSA samples and 83 healthy controls were compared; a total of 341 individuals' samples were included in this comparison that 258 correspond to PSA and 83 to healthy controls (Figure 2B,E). In this comparison, 121 proteins displayed a normal distribution and there are two differential groups of proteins according to the relative abundance level. Regarding high-relative abundance, 53 proteins are in this group, such as SERPIN, COMP, TNF, ILR1, ..., which have been previously reported in these pathologies.[19] On the other side, a set of 68 of these proteins show a low-relative abundance level, which their functions are associated with functions described for the proteins with high-relative abundance proteins, among additional proteins involved in post-translational protein phosphorylation (i.e., SPARCL1, APOA1, ...), which are different than at a high-abundance expression level (Supplementary Table 3).

*3.2.1.3. RA vs Controls***.** At first glance, 82 proteins are differentially observed across 192 RA and 83 healthy controls, where 61 of the 82 displayed a high-relative abundance, which most of them are associated with integrin cell surface signaling (i.e., LGALS3BP, LAMP2, SRGN, APBB1IP0, ...), extracellular matrix organization/degradation (i.e., FCN1, SPARC, NCF2, COL12A1, MMP1, OMD, ...), and innate immune response (i.e., C3, C9, ...) (Figure 2C,F) (Supplementary Table 4). On the other hand, a low abundance profile is also observed (21 proteins) that is involved in metabolic signaling pathways (i.e., GSTO1, MIA3, PLTP, ...), dys-regulation of osteoblast differentiation (i.e., UCMA), components of extracellular matrix (i.e., VIT, COL9, ...) (Supplementary Table 4).

Moreover, in both differential profiles, there are proteins related to immune response (such as TNF, CD53, C3, C9, ...) which has been also highlighted in both databases (PURPOSE and PubPular databases[13]).

3.2.2. Across Pathological Groups (AO, RA, PSA). As the full cohort has been simultaneously screened by the same functional proteomics approach, then comparisons are feasible and also useful to decipher potential biomarkers candidates. In Supplementary Table 4, differential protein profiles (based on relative abundance) are listed for each side-by-side comparison. In general, it is observed that most of the proteins are related to innate immune signaling pathways, extracellular matrix organization/degradation, metabolic signaling pathways, ... Among them, >70% of them have been previously reported in the PURPOSE and PubPular databases.[13]

3.3. Exploring Differential Protein Profiles as Potential Biomarkers Candidates

As several differential protein profiles are observed in the multiple performed comparisons, it is interesting to explore if any of these set of proteins could be potentially a panel of biomarkers useful for diagnosis. In Figure 3, it is observed the presence of multiple proteins profiles which could be used to classified samples; however, it is not clearly distinguished a well-defined protein profile for suitable and optimal classification according to the pathology (for the 3 analyzed pathological groups (PC1 is displaying 81.6% as high correlation level) (Figure 3 and Supplementary Figure 4). Then, it might be expected that it could influence in the contribution of information to any model designed from these data sets.

Bearing in mind that most of the selected proteins belong to similar signaling pathways, it is expected a high correlation between the differential protein profiles observed above in the side-by-side comparisons (Supplementary Figure 5). Then, cluster analysis was performed for the 30 selected proteins against 960 screened samples which results in 4 principal clusters (Figure 4, Supplementary Table 5). Each of these groups of proteins is directly related with one particular cellular function being: (i) cell repair mechanism and inflammatory related proteins (cluster 1): SERPINA1, SLC11A1, NCF2, TNC, LUM, LTBP2, HRG, SOD2, CSPGA, C1S, CRTAC1, HABP2, IL1RAP; (ii) cell signaling pathways (cluster 2): IGFALS, IGFALS, HTRA, TNF, PLTP, SPARCL1, COMP,

IGSF6; (iii) bone generation and regeneration (cluster 3): SRGN, SERPINA4, COL6A1, GC, COL11A2; (iv) cell migration (cluster 4): UCMA, MIA3, SPARC.

3.4. Design a Biomarker Panel for Diagnosis

Once it was identified and selected a set of protein candidates from a differential profile, the next step was to decipher whether these proteins provide any prediction value for further or future analysis. Since most of the analytes belong to the same or similar signaling pathways, there was expected a high correlation between them and subsequently the individual contribution to the model could be decreased with each addition. As it was previously observed, patients cluster the analysis of patient's clusters may be used as a predictor method; however, protein cluster analysis might not be considered as a predictor value. Having in mind that the pathology is variable with more than two categories (controls, osteoarthritis, psoriatic arthritis, rheumatoid arthritis), then a multinomial logistic regression model is proposed. A relationship is established between the pathology (dependent variable) and proteins of each cluster (independent variables) through the following multinomial logistic regression function and Pearson correlation. It is observed (Supplementary Figure 6) that several proteins display an independent correlation with the groups of patients (controls, pathological, ...), with a high-correlation index when used in combination.

Thus, a first model was defined to predict any of the pathologies against healthy controls. Here, a set of 16 proteins of the 30 (the regression model displays a pseudo-R^2 McFadden coefficient 0.8553) allows one to determine the probability of a patient presenting any of the pathologies is 10.4278 (constant) times the probability of not presenting such a pathology, that is, of being healthy (controls) (Figure 5, Supplementary Figure 7)).

Then, several models inside pathologies have been also evaluated and a few protein panels have been proposed as prediction models:

(i) Osteoarthritis (OA) Panel: This set of proteins is SPARC, PLTP, SPARCL1, IGFALS, NCF2, COL11A2, SLC11AA, HTRA, HRG, COMP, IGFALS, CSPG4, which correctly classify 75.1% of the samples (a pseudo-R^2 McFadden coefficient 0.2664) (Supplementary Table 6). The ROC curve is displayed in Figure 5B.

(ii) PSA Panel: In this case, 15 proteins of the 30 previously selected (SLC11A1, NCF2, CSPG4, MIA3.1, GC.1, CSPG4, SOD2, C1S, COMP.1, SERPINA1, SERPINA4, COL11A2, IGFALS, TNC, LTBP2), with 72,63% properly classified (and pseudo-R^2 McFadden coefficient 0.2018) (Supplementary Table 6). In Figure 5C is depicted the ROC curve of this panel of proteins.

(iii) RA Panel: This prediction model is based on 14 proteins (PLTP.1, SPARCL1, UCMA, GC, COMP.1, CSPG, SERPINA4, C1S, SOD2, SERPINA1, IGFALS, HTRA1, TNC, CSPG4) that properly classify 80.96% of the cases (a pseudo-R^2 McFadden coefficient 0.415) (Supplementary Table 6). The ROC curve is displayed in Figure 5D.

In a similar approach, it has been evaluated the prediction capacity inside the pathologies, by a binomial logistic model as reported in Supplementary Figure 8. There is a set of proteins which might be useful to distinguish between OA vs RA, APS vs OA, and APS vs RA (Supplementary Table 7). In all these comparisons, the area under the curve (AUC) for each ROC (0.781, 0.902, 0.869) (Supplementary Figure 8) shows the classification power of each model applied is quite optimal which could be also highly useful to discriminate between pathologies.

Then, multiple combinations of proteins, from the 30 selected panels detected by protein microarrays, allow one to discriminate between healthy vs pathologies and between pathologies with high accuracy and sensitivity. Thus, it is proof that the designed multipronged approach (combining databases, functional proteomics and well-characterized clinical cohort) is useful for biomarker discovery and further validation of these pathologies.

Subsequently, the creation of the binomial logistic regression models is making it possible to verify and confirm which one of the 30 identified proteins is acquiring relevance in each model and its particular precision in each comparison (by means of the percentage of individuals properly classified and the relative frequency criterion (Supplementary Table 6, Supplementary Figure 7)). In general, for all the models established, the AUC was in the range 80–95% for all the group comparisons, and these panels of proteins might be verified as biomarkers in these pathologies in further analysis. Nonetheless, it can be noted that it seems that multiple proteins are only relevant in a specific model for particular comparison (i.e., OA vs healthy) which will require further confirmation studies.

4. CONCLUSIONS

The results showed differential serum protein profiles among patient groups. In fact, a set of proteins showing significantly different levels among groups has been identified. Altered protein profiles in the pathological samples (OA, RA, PSA) were proven substantially higher when compared to healthy control samples. Unsupervised and supervised machine learning approaches allowed the accurate classification of patients based on this set of proteins, which constitute a panel of potential serum biomarkers. Hence, several panels (from 6 to 16 protein) have been identified for each pathology (OA, RA, PSA). In summary, 30 different proteins could be considered as a potential panel to cover all these pathologies. From the functional point of view, these proteins are involved in inflammatory response, lipid metabolism, bone and ECM degradation and/or remodelling, which correlate with ethology and physiology of these pathologies. The synergy between popular proteins (reported in curated published databases available at HPP) and functional proteomics approaches (i.e., antibody microarrays) seems to be a quite useful and powerful strategy in the B/D-HPP initiative to accelerate biomarker discovery and validation pipelines. Nevertheless, further studies will be necessary in order to accurately quantify differences in protein levels, with a larger cohort of patients, including other rheumatic/inflammatory pathologies and exploring the correlation with biological-clinical information (i.e., therapies, disease status, etc.). In addition, this approach may be able to combine with other molecular information generated by “-

omics" studies with imaging and clinical data that might be highly valuable to facilitate precision medicine strategies in OA.

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jproteome.2c00387.

Supplementary Table 1: List of antibodies in these arrays (XLSX)
Supplementary Table 2: Relative protein abundance level in OA vs controls (XLSX)
Supplementary Table 3: Relative protein abundance level in PSA vs controls (XLSX)
Supplementary Table 4: Relative protein abundance level in RA vs controls (XLSX)
Supplementary Table 5: Differential protein profiles of significant proteins in each cluster (PDF)
Supplementary Table 6: Parameters of the logistic multinomial regression (XLSX)
Supplementary Table 7: Differential proteins profiles in each pathological comparison (XLSX)
Supplementary Figures 1–8 (PDF)

Corresponding Authors

Manuel Fuentes - *Department of Medicine and General Cytometry Service-Nucleus, Cancer Research Centre (IBMCC/CSIC/USAL/IBSAL), 37007Salamanca, Spain*; *Proteomics Unit, Cancer Research Centre (IBMCC/CSIC/USAL/IBSAL), 37007Salamanca, Spain*; https://orcid.org/0000-0002-7305-3766; Email: mfuentes@usal.es
Francisco J. Blanco - *Unidad de Proteómica, Grupo de Investigación de Reumatología (GIR), Instituto de Investigación Biomédica de A Coruña (INIBIC), Complexo Hospitalario Universitario de A Coruña (CHUAC), Sergas. C/As Xubias de Arriba 84, 15006A Coruña, Spain*; *Grupo de Investigación de Reumatología y Salud (GIR-S), Departamento de Fisioterapia, Medicina y Ciencias Biomédicas, Centro de investigaciones Avanzadas (CICA),*

Universidade da Coruaña (UDC), 15008A Coruña, Spain; https://orcid.org/0000-0001-9821-7635; Email: Francisco.Blanco.Garcia@sergas.es

Authors

Cristina Ruiz-Romero - *Unidad de Proteómica, Grupo de Investigación de Reumatología (GIR), Instituto de Investigación Biomédica de A Coruña (INIBIC), Complexo Hospitalario Universitario de A Coruña (CHUAC), Sergas. C/As Xubias de Arriba 84, 15006A Coruña, Spain*; *Centro de Investigación Biomédica en Red de Bioingeniería, Biomateriales y Nanomedicina (CIBER-BBN), Av. Monforte de Lemos, 3-5. Pabellón 11, 28029Madrid, Spain*; https://orcid.org/0000-0001-7649-9803

Sara Misiego - *Department of Medicine and General Cytometry Service-Nucleus, Cancer Research Centre (IBMCC/CSIC/USAL/IBSAL), 37007Salamanca, Spain*

Pablo Juanes-Velasco - *Department of Medicine and General Cytometry Service-Nucleus, Cancer Research Centre (IBMCC/CSIC/USAL/IBSAL), 37007Salamanca, Spain*

Alicia Landeira-Viñuela - *Department of Medicine and General Cytometry Service-Nucleus, Cancer Research Centre (IBMCC/CSIC/USAL/IBSAL), 37007Salamanca, Spain*

Adrián Torres-Roda - *Department of Medicine and General Cytometry Service-Nucleus, Cancer Research Centre (IBMCC/CSIC/USAL/IBSAL), 37007Salamanca, Spain*

Héctor Lorenzo-Gil - *Department of Medicine and General Cytometry Service-Nucleus, Cancer Research Centre (IBMCC/CSIC/USAL/IBSAL), 37007Salamanca, Spain*

María González-González - *Department of Medicine and General Cytometry Service-Nucleus, Cancer Research Centre (IBMCC/CSIC/USAL/IBSAL), 37007Salamanca, Spain*; https://orcid.org/0000-0001-8115-1086

Ángela P. Hernández - *Department of Medicine and General Cytometry Service-Nucleus, Cancer Research Centre (IBMCC/CSIC/USAL/IBSAL), 37007Salamanca, Spain*; *Department of Pharmaceutical Sciences: Organic Chemistry, Faculty of Pharmacy, University of Salamanca, CIETUS, IBSAL, 37007Salamanca, Spain*; https://orcid.org/0000-0002-5585-9918

Lucía Lourido - *Unidad de Proteómica, Grupo de Investigación de Reumatología (GIR), Instituto de Investigación Biomédica de A Coruña (INIBIC), Complexo Hospitalario Universitario de A Coruña (CHUAC), Sergas. C/As Xubias de Arriba 84, 15006A Coruña, Spain*; https://orcid.org/0000-0001-9395-0598

Ronald Sjöberg - *Department of Protein Science, SciLifeLab, KTH Royal Institute of Technology, 114 28Stockholm, Sweden*

Elisa Pin - *Department of Protein Science, SciLifeLab, KTH Royal Institute of Technology, 114 28Stockholm, Sweden*
Javier de Las Rivas - *Bioinformatics and Functional Genomics Research Group, Cancer Research Centre (IBMCC/CSIC/USAL/IBSAL), 37007Salamanca, Spain*;
https://orcid.org/0000-0002-0984-9946
José Manuel Sánchez-Santos - *Bioinformatics and Functional Genomics Research Group, Cancer Research Centre (IBMCC/CSIC/USAL/IBSAL), 37007Salamanca, Spain*
Peter Nilsson - *Department of Protein Science, SciLifeLab, KTH Royal Institute of Technology, 114 28Stockholm, Sweden*

Notes

The authors declare no competing financial interest.

Acknowledgments

We gratefully acknowledge financial support from the Spanish Health Institute Carlos III (ISCIII) for the grants FIS PI21/01545, PI19/01206, PI20/00793, CB16/12/00400, and CB06/01/0040. We also acknowledge Fondos FEDER (EU) and Junta Castilla-León (COVID-19 grant COV20EDU/00187). The Proteomics Units belong to ProteoRed, PRB3-ISCIII, supported by grants PT17/0019/0023 and PT17/0019/0014 of the PE I+D+I 2017-2020, funded by ISCIII and FEDER. This research work was funded by the European Commission – NextGeneration EU, through CSIC's Global Health Platform (PTI+ Salud Global). This work was performed in the framework of NanoMedicine CSIC HUB (ref 202180E048). P.J.-V. is supported by JCYL PhD Program "Nos Impulsa-JCYL" and scholarship JCYL-EDU/601/2020. A.L.-V. is supported by VIII Centenario-USAL PhD Program. L.L. is supported by a Sara Borrell contract (CD19/00229).

REFERENCES

(1) Hodgkinson, T.; Kelly, D. C.; Curtin, C. M.; O'Brien, F. J. Mechanosignalling in cartilage: an emerging target for the treatment of osteoarthritis. *Nat. Rev. Rheumatol.* **2022**, *18*, 67– 84, DOI: 10.1038/s41584-021-00724-w

(2) Li, Y. S.; Xie, W. Q.; Xiao, W. F.; Dou, D. Progress in osteoarthritis research by the National Natural Science Foundation of China. *Bone Res.* **2022**, 10, DOI: 10.1038/s41413-022-00207-y .

(3) Lohmander, L. S.; Felson, D. Can we identify a 'high risk'patient profile to determine who will experience rapid progression of osteoarthritis?. *Osteoarthritis Cartilage* **2004**, *12*, 49– 52, DOI: 10.1016/j.joca.2003.09.004

(4) Haseeb, A.; Haqqi, T. M. Immunopathogenesis of osteoarthritis. *Clinical Immunol.* **2013**, *146*, 185– 196, DOI: 10.1016/j.clim.2012.12.011

(5) Hunter, D. J.; Nevitt, M.; Losina, E.; Kraus, V. Biomarkers for osteoarthritis: Current position and steps towards further validation. *Best Practice Res. Clin. Rheumatol.* **2014**, *28*, 61– 71, DOI: 10.1016/j.berh.2014.01.007

(6) Tardif, G.; Pare, F.; Gotti, C.; Roux-Dalvai, F.; Droit, A.; Zhai, G.; Sun, G.; Fahmi, H.; Pelletier, J. P.; Martel-Pelletier, J. Mass spectrometry-based proteomics identify novel serum osteoarthritis biomarkers. *Arthritis Res. Therapy* **2022**, 24, DOI: 10.1186/s13075-022-02801-1 .

(7) Ruiz-Romero, C.; Blanco, F. J. Proteomics role in the search for improved diagnosis, prognosis and treatment of osteoarthritis. *Osteoarthritis Cartilage* **2010**, *18*, 500– 509, DOI: 10.1016/j.joca.2009.11.012

(8) Ruiz-Romero, C.; Calamia, V.; Albar, J. P.; Casal, J. I.; Corrales, F. J.; Fernandez-Puente, P.; Gil, C.; Mateos, J.; Vivanco, F.; Blanco, F. J. The Spanish biology/disease initiative within the human proteome project: Application to rheumatic diseases. *J. Proteomics* **2015**, *127*, 406– 413, DOI: 10.1016/j.jprot.2015.01.023

(9) Thurlings, R. M.; Tas, S. W.; Tak, P. P. Rheumatoid Arthritis and Other Inflammatory Articular Diseases. *Inflammation - From Molecular and Cellular Mechanisms to the Clinic* **2017**, 1105– 1140, DOI: 10.1002/9783527692156.ch42

(10) Van Eyk, J. E.; Corrales, F. J.; Aebersold, R.; Cerciello, F.; Deutsch, E. W.; Roncada, P.; Sanchez, J. C.; Yamamoto, T.; Yang, P.; Zhang, H.Highlights of the Biology and Disease-driven Human Proteome Project, 2015–2016. *J. Proteome Res.* **2016**, *15*, 3979–3987, DOI: 10.1021/acs.jproteome.6b00444

(11) Lam, M. P. Y.; Venkatraman, V.; Xing, Y.; Lau, E.; Cao, Q.; Ng, D. C. M.; Su, A. I.; Ge, J. B.; Van Eyk, J. E.; Ping, P. P. Data-Driven Approach To Determine Popular Proteins for Targeted Proteomics Translation of Six Organ Systems. *J. Proteome Res.* **2016**, *15*, 4126– 4134, DOI: 10.1021/acs.jproteome.6b00095

(12) Aebersold, R.; Bader, G. D.; Edwards, A. M.; van Eyk, J. E.; Kussmann, M.; Qin, J.; Omenn, G. S. The Biology/Disease-driven Human Proteome Project (B/D-HPP): Enabling Protein Research for the Life Sciences Community. *J. Proteome Res.* **2013**, *12*, 23– 27, DOI: 10.1021/pr301151m

(13) Ruiz-Romero, C.; Lam, M. P. Y.; Nilsson, P.; Onnerfjord, P.; Utz, P. J.; Van Eyk, J. E.; Venkatraman, V.; Fert-Bober, J.; Watt, F. E.; Blanco, F. J. Mining the Proteome Associated with Rheumatic and Autoimmune Diseases. *J. Proteome Res.* **2019**, *18*, 4231– 4239, DOI: 10.1021/acs.jproteome.9b00360

(14) 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*, 4267– 4278, DOI: 10.1021/acs.jproteome.8b00393

(15) Yu, K. H.; Lee, T. L. M.; Chen, Y. J.; Re, C.; Kou, S. C.; Chiang, J. H.; Snyder, M.; Kohane, I. S. A Cloud-Based Metabolite and Chemical Prioritization System for the Biology/Disease-Driven Human Proteome Project. *J. Proteome Res.* **2018**, *17*, 4345– 4357, DOI: 10.1021/acs.jproteome.8b00378

(16) Jourquin, J.; Duncan, D.; Shi, Z.; Zhang, B. GLAD4U: deriving and prioritizing gene lists from PubMed literature. *Bmc Genomics* **2012**, 13, DOI: 10.1186/1471-2164-13-S8-S20 .

(17) Tsuruoka, Y.; Miwa, M.; Hamamoto, K.; Tsujii, J.; Ananiadou, S. Discovering and visualizing indirect associations between biomedical concepts. *Bioinformatics* **2011**, *27*, I111– I119, DOI: 10.1093/bioinformatics/btr214

(18) Henjes, F.; Lourido, L.; Ruiz-Romero, C.; Fernandez-Tajes, J.; Schwenk, J. M.; Gonzalez-Gonzalez, M.; Blanco, F. J.; Nilsson, P.; Fuentes, M. Analysis of Autoantibody Profiles in Osteoarthritis Using Comprehensive Protein Array Concepts. *J. Proteome Res.* **2014**, *13*, 5218– 5229, DOI: 10.1021/pr500775a

(19) Sierra-Sánchez, A.l.; Garrido-Martín, D.; Lourido, L.; Gonzalez-Gonzalez, M.; Díez, P.; Ruiz-Romero, C.; Sjöber, R.; Droste, C.; De Las Rivas, J.; Nilsson, P.Screening and validation of novel biomarkers in osteoarticular pathologies by comprehensive combination of protein array technologies. *J. Proteome Res.* **2017**, *16*, 1890– 1899, DOI: 10.1021/acs.jproteome.6b00980

(20) Altman, R.; Asch, E.; Bloch, D.; Bole, G.; Borenstein, D.; Brandt, K.; Christy, W.; Cooke, T. D.; Greenwald, R.; Hochberg, M.Development of criteria for the classification and reporting of osteoarthritis. Classification of osteoarthritis of the knee. Diagnostic and Therapeutic Criteria Committee of the American Rheumatism Association. *Arthritis Rheumatism* **1986**, *29*, 1039– 1049, DOI: 10.1002/art.1780290816

(21) Gonzalez-Gonzalez, M.; Bartolome, R.; Jara-Acevedo, R.; Casado-Vela, J.; Dasilva, N.; Matarraz, S.; Garcia, J.; Alcazar, J. A.; Sayagues, J. M.; Orfao, A.Evaluation of homo-

and hetero-functionally activated glass surfaces for optimized antibody arrays. *Anal. Biochem.* **2014**, *450*, 37– 45, DOI: 10.1016/j.ab.2014.01.002

(22) Landeira-Vinuela, A.; Diez, P.; Juanes-Velasco, P.; Lecrevisse, Q.; Orfao, A.; de las Rivas, J.; Fuentes, M. Deepening into Intracellular Signaling Landscape through Integrative Spatial Proteomics and Transcriptomics in a Lymphoma Model. *Biomolecules* **2021**, *11*, 1776, DOI: 10.3390/biom11121776

(23) Haggmark, A.; Neiman, M.; Drobin, K.; Zwahlen, M.; Uhlen, M.; Nilsson, P.; Schwenk, J. M. Classification of protein profiles from antibody microarrays using heat and detergent treatment. *New Biotechnol.* **2012**, *29*, 564– 570, DOI: 10.1016/j.nbt.2011.10.005

(24) Juanes-Velasco, P.; Landeira-Vinuela, A.; Hernandez, A.-P.; Fuentes, M. Systematic and Rational Design of Protein Arrays in Noncontact Printers: Pipeline and Critical Aspects. *Methods in Molecular Biology (Clifton, N.J.)* **2021**, *2344*, 9– 29, DOI: 10.1007/978-1-0716-1562-1_2

(25) Hawkins, D. M. *Identification of outliers*; Springer: 1980; Vol. 11.

(26) Grimaldi, R. P. *Discrete and Combinatorial Mathematics, 5/e*; Pearson Education: India, 2006.

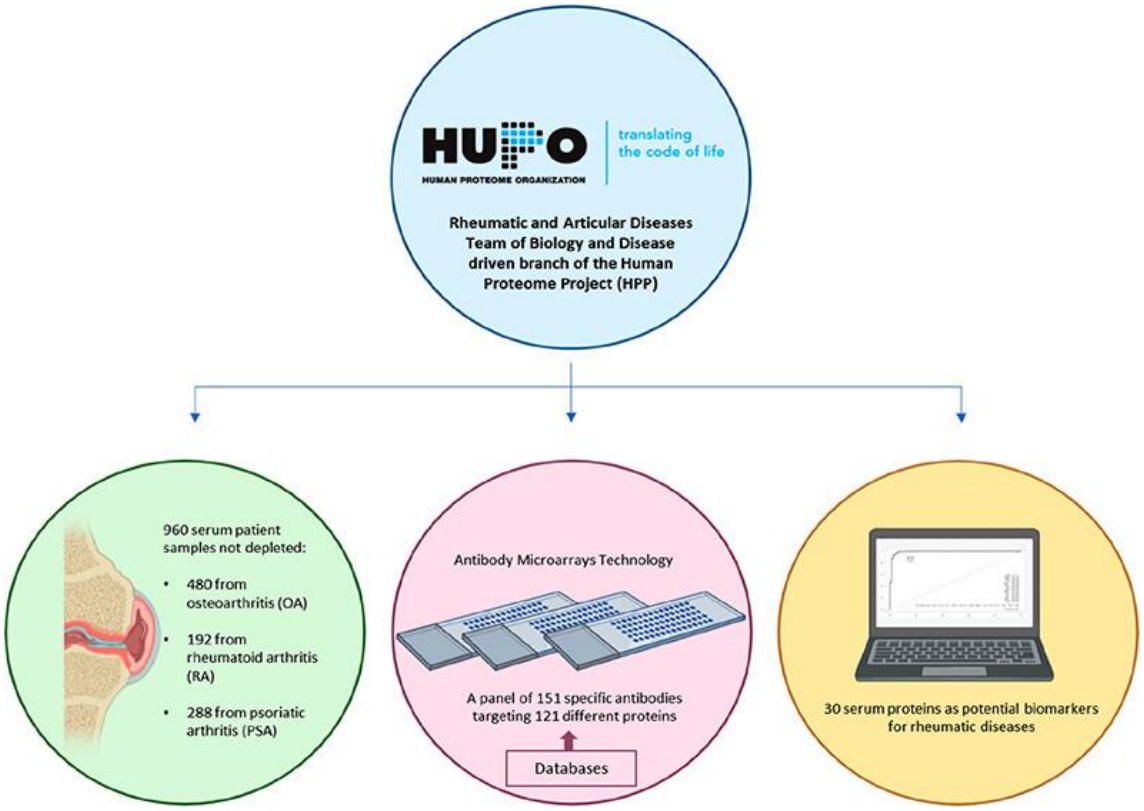
(27) Moore, L. M. *The basic practice of statistics*; W. H. Freeman, 1996.

(28) Sherman, B. T.; Hao, M.; Qiu, J.; Jiao, X. L.; Baseler, M. W.; Lane, H. C.; Imamichi, T.; Chang, W. Z. DAVID: a web server for functional enrichment analysis and functional annotation of gene lists (2021 update). *Nucleic Acids Res.* **2022**, *50*, W216, DOI: 10.1093/nar/gkac194

(29) Shannon, P.; Markiel, A.; Ozier, O.; Baliga, N. S.; Wang, J. T.; Ramage, D.; Amin, N.; Schwikowski, B.; Ideker, T. Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome Res.* **2003**, *13*, 2498– 2504, DOI: 10.1101/gr.1239303

(30) Gonzalez-Gonzalez, M.; Sayagues, J. M.; Munoz-Bellvis, L.; Pedreira, C. E.; de Campos, M. L. R.; Garcia, J.; Alcazar, J. A.; Braz, P. F.; Galves, B. L.; Gonzalez, L. M.; Tracking the Antibody Immunome in Sporadic Colorectal Cancer by Using Antigen Self-Assembled Protein Arrays. *Cancers* **2021**, 13, 2718, DOI: 10.3390/cancers13112718 .

(31) Farrugia, M.; Baron, B. The role of TNF-alpha in rheumatoid arthritis: a focus on regulatory T cells. *J. Clin. Transl. Res.* **2016**, *2*, 84– 90, DOI: 10.18053/jctres.02.201603.005



Abstract

Table 1. Distribution of Analyzed Serum Samples in the Study

Samples (*n*)	Absolute Frequency	Relative Frequency
Controls	83	0.08
Osteoarthritis (OA)	427	0.44
Rheumatoid Arthritis (RA)	258	0.26
Psoriatic Arthritis (PSA)	192	0.20
Total (*n*)	960	1

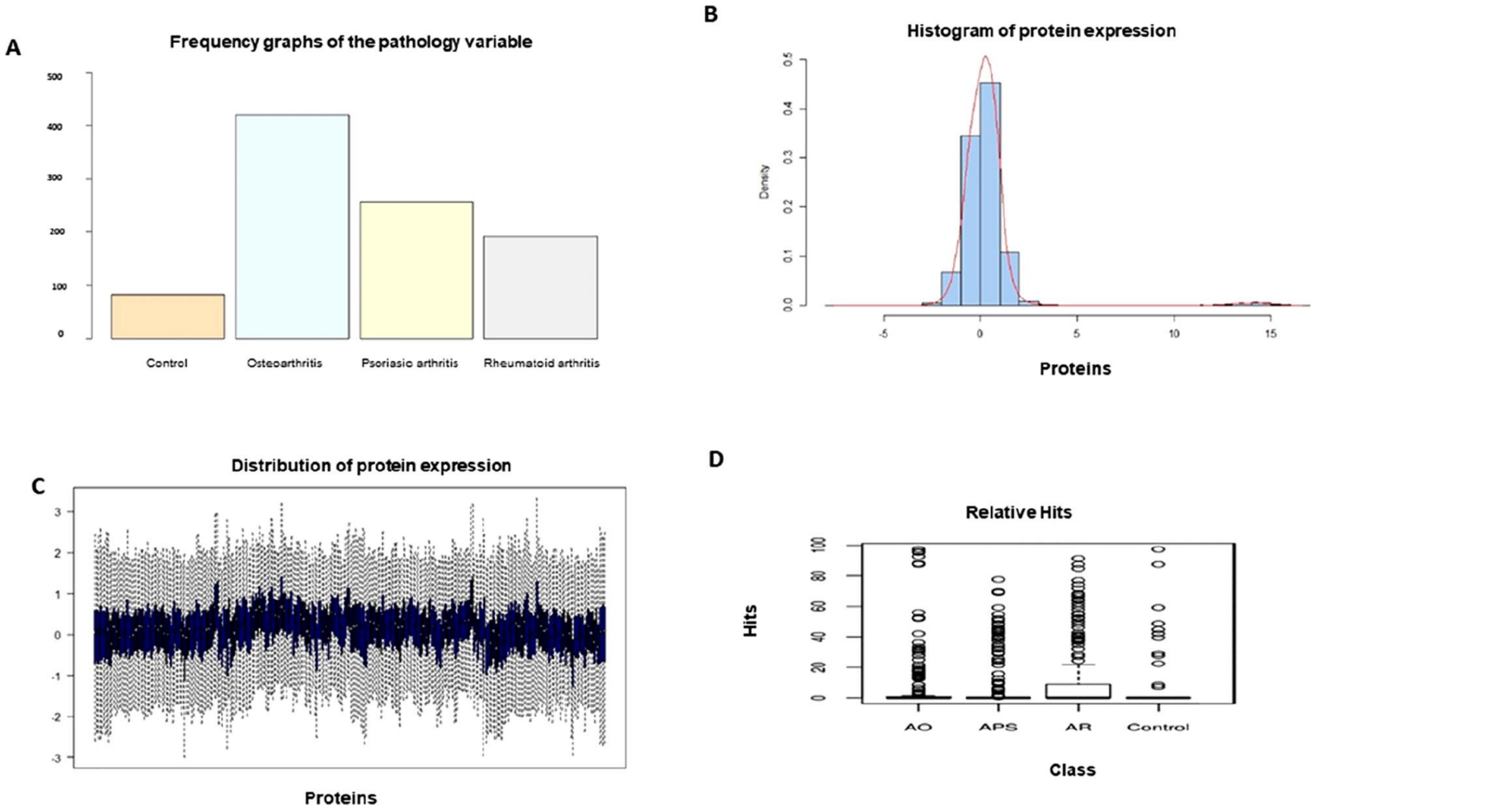


Figure 1. Summary of array performance. Panel A: Frequency graphs of the pathology. Panel B: Histogram of protein expression. Panel C: Distribution of protein expression. Panel D: Relative number of hits by pathological groups.

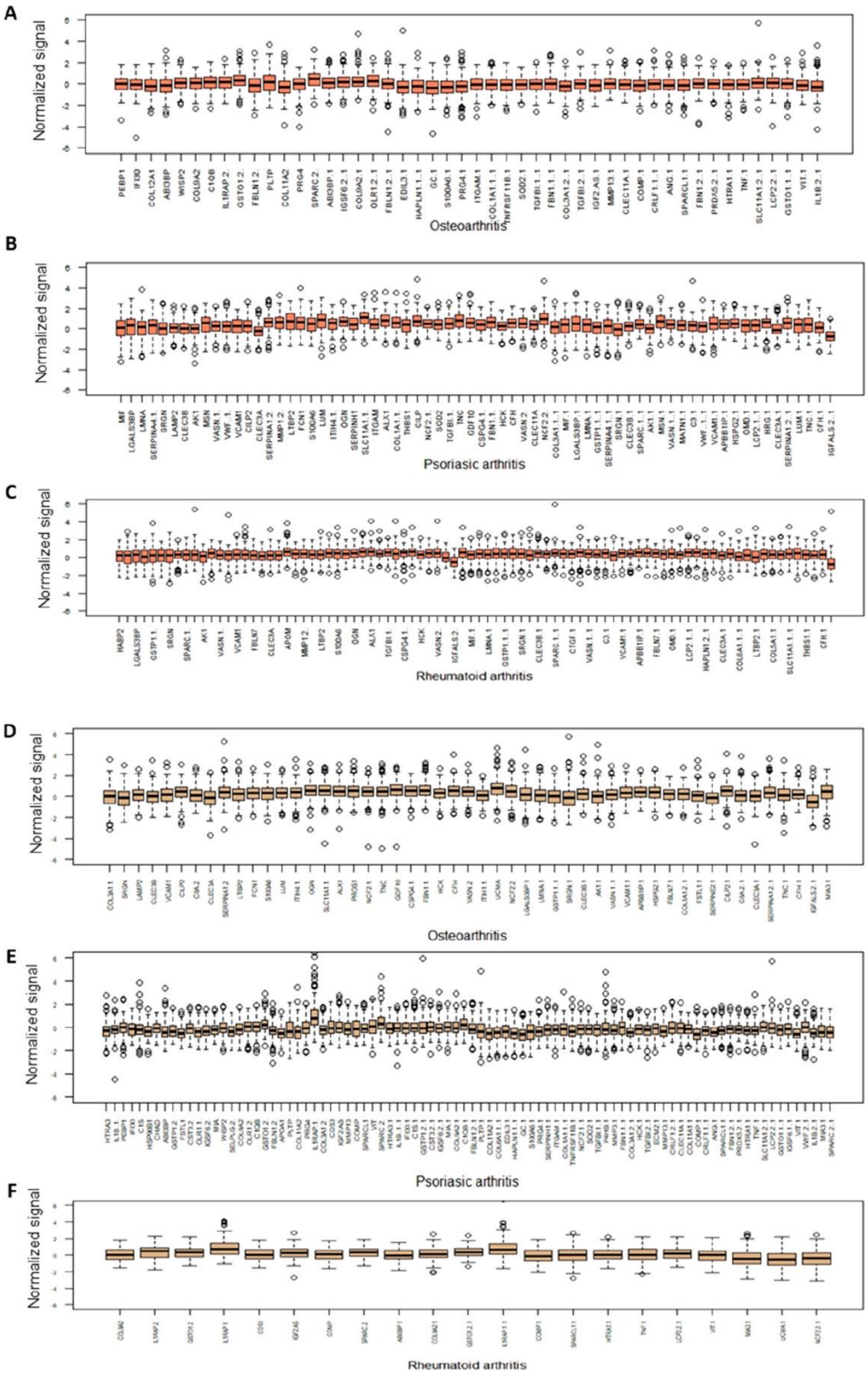


Figure 2. Boxplots of up-regulated/down-regulated proteins relative abundance in all the comparisons. Panel A: Up in OA. Panel B: Up in PSA. Panel C: Up in RA. Panel D: Down in OA. Panel E: Down in PSA. Panel F: Down in RA.

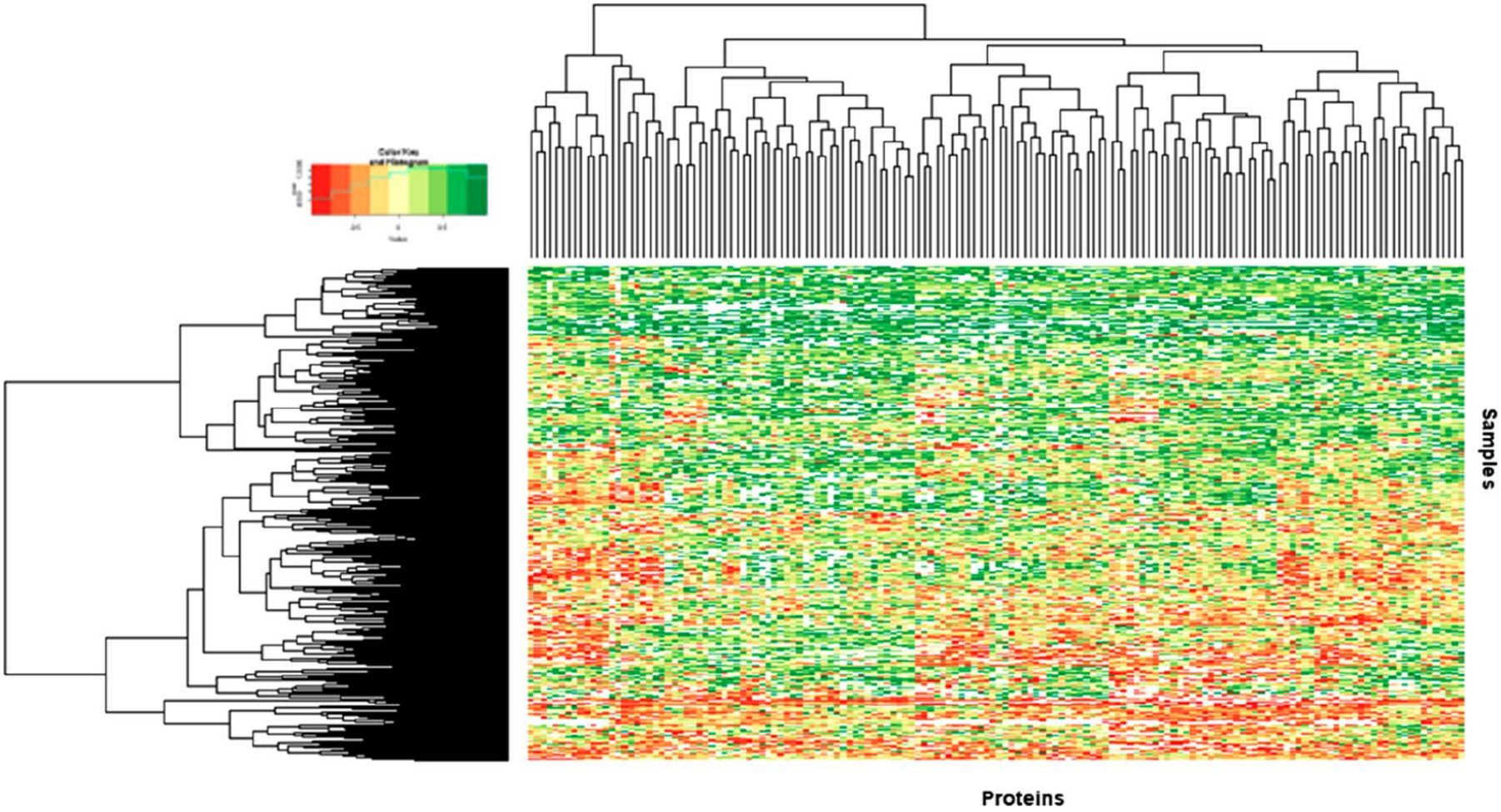


Figure 3. Global differential protein profile. Cluster analysis of all patients serum samples by affinity proteomics.

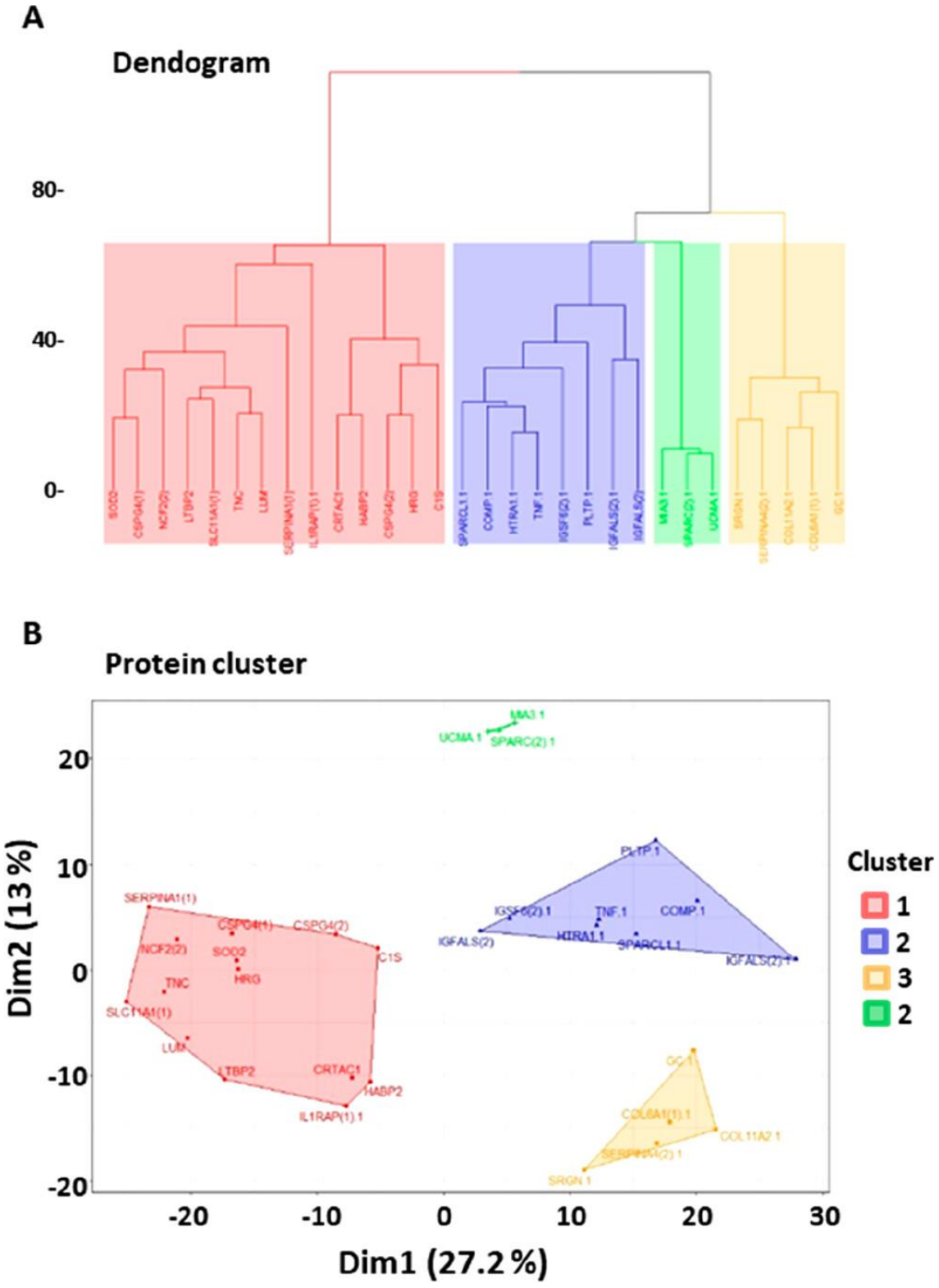


Figure 4. Exploring differential protein profiles. Panel A: Dendogram describing minimal number of clusters according to protein profiles. Panel B: PCA analysis and grouping sample distribution.

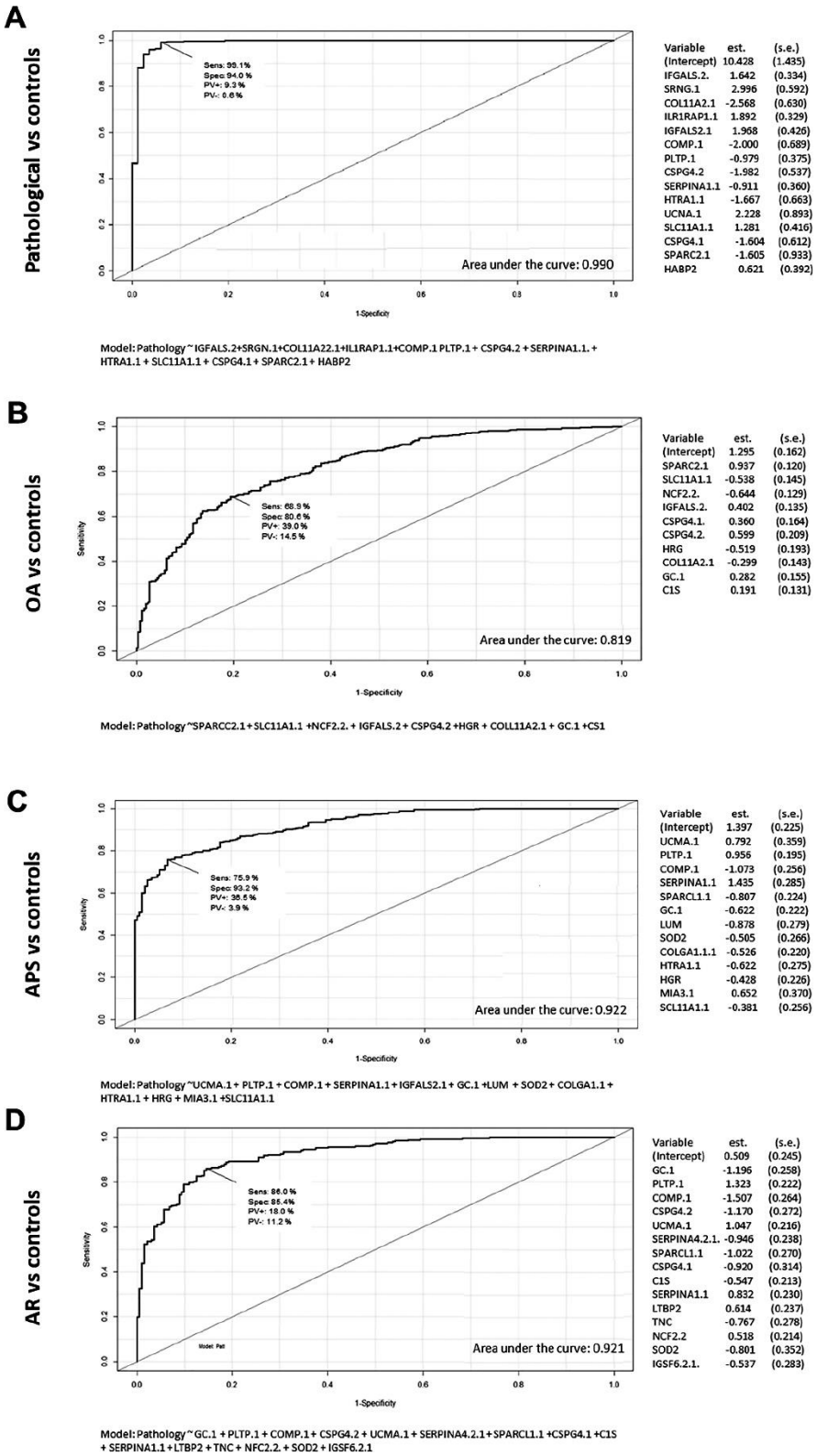


Figure 5. Evaluation of differential serum protein profiles as potential biomarkers panel for diagnosis. Panel A: ROC curve for pathological vs controls. Panel B: ROC curve for OA vs controls. Panel C: ROC curve for APS vs controls. Panel D: ROC curve for RA vs controls.