+ MODEL

Genes & Diseases (2024) 11, 101016

Available online at www.sciencedirect.com

ScienceDirect

journal homepage: www.keaipublishing.com/en/journals/genes-diseases

RAPID COMMUNICATION

Potential risk genes for primary Sjogren's syndrome from a meta-analysis by linear regression and random forest classification

Primary Sjögren's syndrome (pSS) is one of the most widespread autoimmune diseases with unknown origin, characterized by a lymphocytic infiltrate of the exocrine glands and the production of anti-SSA/Ro and anti-SSB/La antibodies that cause dysfunction and destruction mainly of salivary and lachrymal glands, leading to dry eyes and dry mouth. In the absence of a standardized evidence-based screening tool to decide which patients with dry eye must be referred for study of this pathology, there is a tendency to have a continuous underdiagnosis of the disease, causing the calculation of its prevalence to be inaccurate.[1] To improve our understanding of the underlying molecular nature, we analyzed public microarray-based gene expression profiles of salivary glands (GSE23117) and saliva (GSE7451) from primary Sjögren's syndrome (pSS) patients and controls. Unlike the two microarrays published before, we showed a specific common set of genes and an extensive study of biological pathways and networks showing a potential detonating factor. We included a machine learning model to distinguish pSS patients from controls, opening a possible means of detecting biomarkers to diagnose, monitor the response to treatment, and predict the prognosis of pSS.

Through linear regression, we first identified differentially expressed genes (DEGs) between pSS patients and controls within each dataset. For the GSE23117 dataset, 1561 significant genes ($P < 0.05$; FDR corrected) were found, of which 860 were overexpressed and 701 were underexpressed in pSS patients and the five more significant genes overexpressed in this study group were *GBP1*, *CXCL9*, *GBP3*, *NLRC5*, and *CXCL10* (Fig. S1A and Table S1). For the GSE7451 dataset, 7 significant genes ($P < 0.05$; FDR corrected) were found overexpressed in pSS patients, where *IFIT3*, *ISG15*, *IFIT1*, *RSAD2*, and *CMPK2* were the 5 more significant genes (Fig. S1B and Table S1). The similarity between significant gene expression profiles of different samples was graphed in a heatmap with an unsupervised bidirectional hierarchical cluster analysis (Fig. 1A, B). The heatmap showed a clearly different expression profile of genes between pSS patients and controls. A Fisher's exact test corrected by BH was conducted to assess the strength of the association between both gene lists, founding 5 common genes: *CARD16*, *CMPK2*, *IFIT2*, *IFIT3*, and *RSAD2* ($P = 1.8e{-}05$, odds ratio: 37.7).

Interestingly, *IFIT2* and *IFIT3* genes are highly expressed to help *IFIT1* during the intrinsic immune response against viruses such as papillomavirus and hepatitis.[2] Furthermore, *RSDA2* has a double function because it directly suppresses viral replication and facilitates the production of IFN-I mediated by TLR7 and TLR9 which induces transcription of the IFIT gene family to which *IFIT2* and *IFIT3* belong.[3] Regarding CMPK2, it has been involved in the response against various viral diseases and is also found on chromosome 2, adjacent and inverted with respect to RSAD2.[4] Lastly, it should be noted that the molecular function of CARD16 remains unknown and is not yet very clear. One of the described functions of CARD16 is that, when is oligomerized, it promotes CARD-mediated molecular assembly and activation of CASP1, which allows the release of interleukin (IL)-1β, thus initiating the inflammation process.[5] This agrees with the fact that such a gene is overexpressed in pSS patients in this study since one of the main characteristics of this disease is a chronic inflammation of salivary glands.

To know which metabolic pathways were implicated in these genes differentially expressed between the two study groups and which diseases they were related to, we conducted a deep functional analysis. The functional enrichment analysis based on Gene Ontology (GO) terms showed four significant terms for GSE23117, all belonging to the category of biological process (BP), and 80 significant terms

Peer review under responsibility of Chongqing Medical University.

https://doi.org/10.1016/j.gendis.2023.05.015

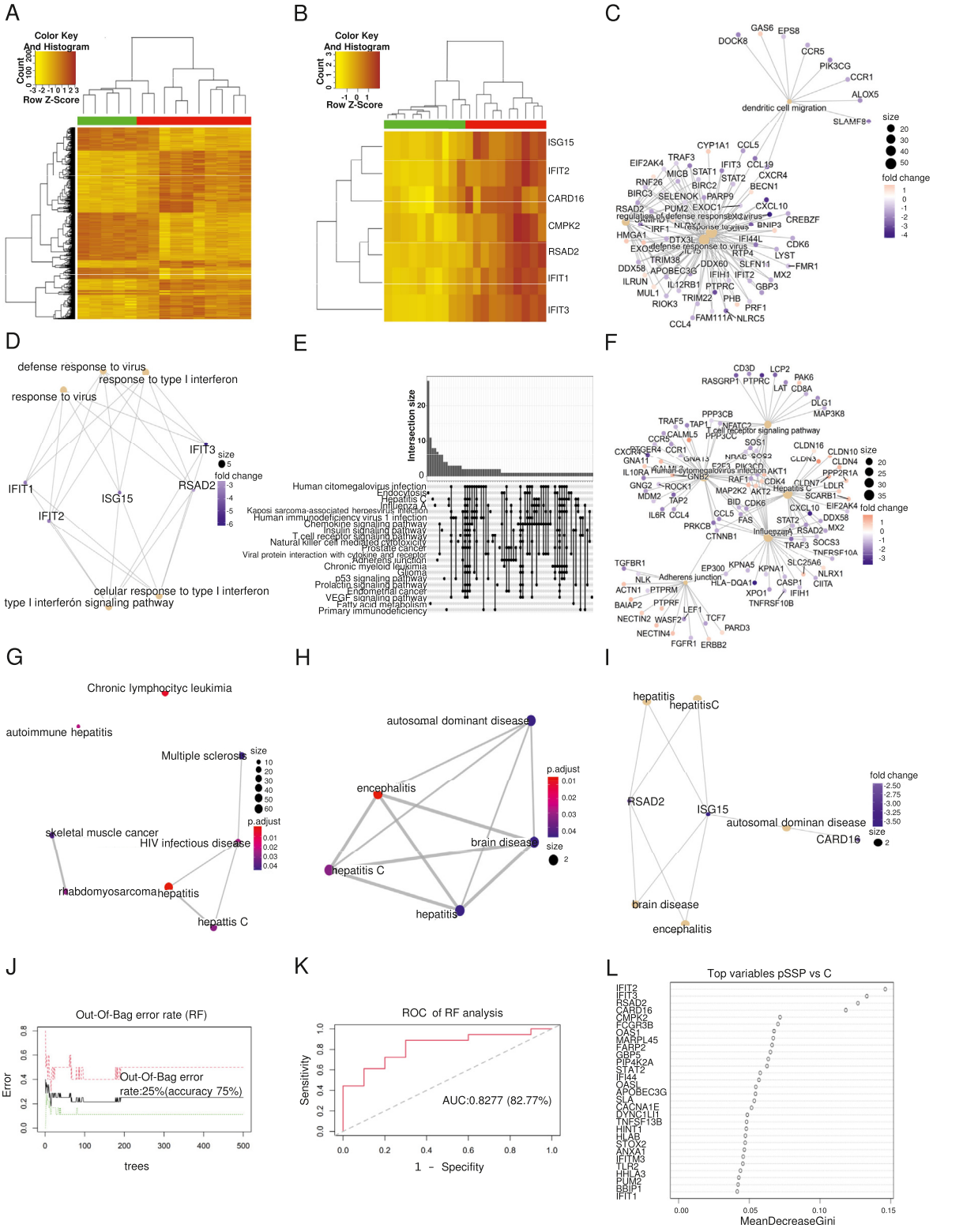

Figure 1 Genetic, functional, and random forest analysis (RF) of this study. **(A, B)** Unsupervised hierarchical clustering analysis of the differentially expressed genes in GSE23117 and GSE7451, respectively. The green color represents controls and the red color represents patients. **(C, D)** Gene-Concept Network of GO term in GSE23117 and GSE7451, respectively. **(E)** The UpSet plot shows the ordered enrichment KEGG pathways of GSE23117. **(F)** Gene-Concept Network of KEGG terms between controls and primary Sjögren's syndrome (pSS) patients in GSE23117. **(G, H)** Network of DO terms enrichment analysis of GSE23117 and GSE7451, respectively. **(I)** Gene-Concept Network of DO terms between control and pSS patients in GSE7451. **(J)** The out-of-bag (OOB) error rate was 25% using the RF machine learning model. **(K)** Receiver operator characteristic (ROC) curve analysis for validating the model, providing an area under the curve (AUC) of 0.8277. **(L)** The top variables contributing to the RF model are shown. The mean decrease in Gini measures the importance of each variable to the model; a higher score indicates a higher importance of the variable.

for GSE7451, 68 in the category of biological process (BP) and 12 in the category of molecular function (MF), being "GO:0051607" ("virus defense response") and "GO:0009615" ("virus response") overrepresented in both datasets (Fig. S2 and Table S2). To evaluate which genes were involved in those enriched pathways and which genes belonged to multiple annotation categories, the top five significant GO terms and the implied genes were represented in a genetic concept network, observing an influence of *IFIT2*, *IFIT3*, and *RSAD2* in response to the virus within pSS patients in both datasets (Fig. 1C, D). On the other hand, the result of the KEGG terms analysis showed a total of 21 significant KEGG terms for GSE23117, all related to viruses, cancer, and processes of the immune system, which shared many genes (Fig. 1E and Table S3). These results were represented in a genetic concept network, where several genes were shared by annotations related to the virus, such as *CTNNB1*, *CXCL9*, *CXCL10*, or *RSAD2*, among others (Fig. 1F). In GSE7451, only one significant KEGG term (hsa05160 -hepatitis C term) was found (Table S3), which was also overrepresented in GSE23117 dataset. Interestingly, the Disease Ontology (DO) terms showed eight significant terms for GSE23117 and five for GSE7451 results (Table S4). The two sample sets showed terms related to hepatitis, as well as some terms associated with other viruses and cancer. These enriched terms were organized into a network with edges connecting overlapping gene sets, identifying the different functional modules (Fig. 1G, H). In the case of GSE23117, only DO terms related to the virus were connected; however, in GSE7451, all DO terms were interconnected. In addition, GSE7451 showed the term "autosomal dominant disease", which was associated with CARD16 (Fig. 1I).

These results seem to coincide with other studies, which point to viruses as triggers of this autoimmune disease and where more than 90% of adults with certain autoimmune diseases present IgG against certain viruses such as Epstein Barr virus, cytomegalovirus, or hepatitis C, among others.[1]

Finally, using a random forest (RF) machine learning approach, 6230 common genes from the GSE23117 and GSE7451 datasets were analyzed (Fig. 1J). The RF model distinguished pSS patients from controls with an 82.77% prediction accuracy (Fig. 1K), supporting a distinctive genetic profile in pSS patients. The variables contributing most to the model were *IFIT2*, *IFIT3*, *RSAD2*, and *CARD16*, followed by *CMPK2* (Fig. 1L). These results were confirmed when the expression of the top four individual genes differentiating patients from controls was assessed by conventional statistical analysis (Fig. S3A) in the global dataset. Notably, receiver operator characteristic (ROC) analysis of these genes showed an AUC of 0.883, 0.973, 0.917, and 0.927, respectively (Fig. S3B).

In conclusion, upregulated genes in pSS patients had previously been pointed out as potentially related to this disease in previous literature, proposing IFIT2, IFIT3, RSAD2, CARD16, and CMPK2 expression as possible biological biomarkers for pSS diagnosis. Furthermore, there are no published studies in the literature on CARD16 linked with pSS in salivary glands and saliva samples, so we raise the importance of this gene in the development of this disease. The results of the functional enrichment analysis of this study propose pathways related to the hepatitis C virus as a possible trigger for pSS disease, fitting the results of other studies in the literature. Lastly, few studies have addressed machine learning in this disease and although our results are based on a small cohort, we propose a novel possible classification model of this disease in which the genes with the most weight coincide with those found using Fisher's exact test corrected by BH between both gene lists.

Author contributions

TC and TTM planned and designed the study. TC conducted the search of datasets and data analysis and both authors reviewed the results. TC wrote the manuscript and both authors read and approved the manuscript.

Conflict of interests

Both authors declare no conflict of interests.

Funding

Tomás Cerdó was supported by 'Sara Borrell' programmes (CD21/00187) from the Institute of Health Carlos III (ISCIII).

Data availability

The two datasets analyzed during the current study are available in the Gene Expression Omnibus (GEO) repository (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE7451; https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE23117).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.gendis.2023.05.015.

References

1. Liu Z, Chu A. Sjögren's syndrome and viral infections. *Rheumatol Ther*. 2021;8(3):1051–1059.
2. Fleith RC, Mears HV, Leong XY, et al. IFIT3 and IFIT2/3 promote IFIT1-mediated translation inhibition by enhancing binding to non-self RNA. *Nucleic Acids Res*. 2018;46(10):5269–5285.
3. Jang JS, Lee JH, Jung NC, et al. Rsad2 is necessary for mouse dendritic cell maturation via the IRF7-mediated signaling pathway. *Cell Death Dis*. 2018;9(8):823.
4. El-Diwany R, Soliman M, Sugawara S, et al. CMPK2 and BCL-G are associated with type 1 interferon-induced HIV restriction in humans. *Sci Adv*. 2018;4(8), eaat0843.
5. Karasawa T, Kawashima A, Usui F, et al. Oligomerized CARD16 promotes caspase-1 assembly and IL-1β processing. *FEBS Open Bio*. 2015;5:348–356.

Tomás Cerdó[a,b,*], Teresa Torres Moral[c,d,e,**]

[a] *Maimonides Biomedical Research Institute of Cordoba (IMIBIC), Reina Sofia University Hospital, University of Cordoba, Córdoba 14004, Spain*
[b] *Centre for Rheumatology Research, Division of Medicine, University College London, London W1T 4JF, UK*
[c] *Faculty of Computer Sciences, Multimedia and Telecommunication, Universitat Oberta de Catalunya, Barcelona 08018, Spain*
[d] *Dermatology Department, Melanoma Unit, August Pi i Sunyer Biomedical Research Institute (IDIBAPS), and Hospital Clínic, Barcelona 08036, Spain*
[e] *Center for Networked Biomedical Research on Rare Diseases (CIBERER), Carlos III Health Institute, Madrid 28029, Spain*

*Corresponding author. Maimonides Biomedical Research Institute of Cordoba (IMIBIC), Reina Sofia University Hospital, University of Cordoba, Córdoba 14004, Spain

**Corresponding author. Faculty of Computer Sciences, Multimedia and Telecommunication, Universitat Oberta de Catalunya, Barcelona 08018, Spain
E-mail addresses: tomas.craez@imibic.org (T. Cerdó), ttorresmo@uoc.edu (T. Torres Moral)

25 December 2022
Available online 4 July 2023