Data Descriptor

Single-Nucleotide Variants in PADI2 and PADI4 and Ancestry Informative Markers in Interstitial Lung Disease and Rheumatoid Arthritis among a Mexican Mestizo Population

Karol J. Nava-Quiroz , Jorge Rojas-Serrano , Gloria Pérez-Rubio , Ivette Buendia-Roldan , Mayra Mejía , Juan Carlos Fernández-López , Espiridión Ramos-Martínez , Luis A. López-Flores , Alma D. Del Ángel-Pablo , Ramcés Falfán-Valencia *

Posted Date: 27 September 2023

doi: 10.20944/preprints202309.1802.v1

Keywords: PAD4/PADI4; PAD2/PADI2; Interstitial Lung Disease; Rheumatoid Arthritis; RA-ILD, AIM

Data Descriptor

Single-Nucleotide Variants in *PADI2* and *PADI4* and Ancestry Informative Markers in Interstitial Lung Disease and Rheumatoid Arthritis among a Mexican Mestizo Population

Karol J. Nava-Quiroz [1,2], Jorge Rojas-Serrano [3], Gloria Pérez-Rubio [1], Ivette Buendía-Roldán [4], Mayra Mejía [5], Juan Carlos Fernández López [6], Espiridión Ramos-Martínez [7], Luis Alberto López-Flores [1], Alma D. Del Ángel-Pablo [1] and Ramcés Falfán-Valencia [1,*]

1 HLA Laboratory, Instituto Nacional de Enfermedades Respiratorias Ismael Cosío Villegas. Tlalpan Mexico City, 14080. Mexico. krolnava@hotmail.com; glofos@yahoo.com.mx; llopezf92@gmail.com; alyde_08@hotmail.com; rfalfanv@iner.gob.mx
2 Universidad Nacional Autónoma de México (UNAM), Programa de Doctorado en Ciencias Médicas, Odontológicas y de la Salud, Investigación Clínica Experimental en Salud, Bioquímica Clínica.
3 Rheumatology Clinic, Instituto Nacional de Enfermedades Respiratorias Ismael Cosío Villegas. Tlalpan Mexico City, 14080. Mexico. jorroser@gmail.com
4 Translational Research Laboratory on Aging and Pulmonary Fibrosis, Instituto Nacional de Enfermedades Respiratorias Ismael Cosío Villegas. Tlalpan Mexico City, 14080. Mexico, ivettebu@yahoo.com.mx
5 Diffuse Interstitial Lung Disease Clinic, Instituto Nacional de Enfermedades Respiratorias Ismael Cosío Villegas. Tlalpan Mexico City, 14080. Mexico. medithmejia1965@gmail.com
6 Consorcio de Genómica Computacional, Instituto Nacional de Medicina Genómica (INMEGEN). Tlalpan Mexico City, 14610. Mexico. jfernandez@inmegen.gob.mx
7 Experimental Medicine Research Unit, Facultad de Medicina, Universidad Nacional Autónoma de México, Mexico City 06720 México, espiri77mx@yahoo.com
* Correspondence: rfalfanv@iner.gob.mx

Abstract: Rheumatoid arthritis (RA) is an autoimmune disease mainly characterized by joint inflammation. It presents extra-articular manifestations, with the lungs one of the affected areas. Among these, damage to the pulmonary interstitium (Interstitial Lung Disease -ILD) has been linked to proteins involved in the inflammatory process and related to extracellular matrix deposition and lung fibrosis establishment. Peptidyl arginine deiminase enzymes (PAD), which carry out protein citrullination, play a role in this context. A genetic association analysis was conducted on genes encoding two PAD isoforms, PAD2 and PAD4. This analysis also included ancestry informative markers and protein level determination in samples from patients with rheumatoid arthritis, RA-associated ILD, and clinically healthy controls. Significant single nucleotide variations (SNV) and a haplotype were identified as susceptibility factors for ILD-RA development. Elevated levels of PAD4 were found in ILD-RA cases, while *PADI2* showed an association with RA susceptibility. This document presents the data obtained from the conducted research, which has been published.

Dataset: DOI 10.3390/cells12182235 (published manuscript). The datasets generated for this study can be found in ClinVar accessions SCV001422427 – SCV001422434

Keywords: PAD4/*PADI4*; PAD2/*PADI2*; interstitial lung disease; rheumatoid arthritis; RA-ILD; AIM

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 September 2023 doi:10.20944/preprints202309.1802.v1

1. Summary (required)

Rheumatoid arthritis (RA) is an autoimmune disease, primarily inflammatory, affecting the joints [1,2]. Extra-articular manifestations have been described in this condition, with the lung being one of the affected sites. Damage to the pulmonary interstitium (interstitial lung disease ILD) [3]. It has been linked to the involvement of proteins participating in the inflammatory process and associated with extracellular matrix deposition and fibrosis establishment in the lungs, such as peptidyl arginine deiminases (PAD) enzymes that carry out protein citrullination.

The prevalence of musculoskeletal diseases in Mexico has been reported to vary by geographic location [4], leading to an analysis of private markers within the Mexican population obtained by Silva et al. in 2009 [5].

Regarding factors associated with developing both diseases (RA and RA-ILD), smoking has been described as one of the main risk factors, in addition to occupational and environmental factors [6].

2. Data Description

The data obtained in each analysis are presented and described in the tables and figures.

2.1. Study groups

The included groups consisted of patients with rheumatoid arthritis (RA), RA associated with diffuse interstitial lung disease (RA-ILD), as well as a group of clinically healthy subjects (CHS). These groups are displayed in the result tables.

2.2. Genotyping

Table 1 displays the analysis variable at the top of each column, including the gene, the SNV with their respective genotypes or alleles, and the patient or control group. The rows represent the frequency in percentage.

Table 1. Analysis of genotypes and alleles of SNV in *PADI2* and *PADI4*.

		RA-ILD (n= 118)	***RA (n= 133)***	***CHS (n= 616)***
Gene	***Genotype or allele***	**F%**	**F%**	**F%**
PADI2	*rs2057094*			
	GG	35.59	40.54	43.67
	GA	23.73	22.52	30.19
	AA	40.68	36.94	26.14
	G	47.46	51.8	58.77
	A	52.54	48.2	41.23
	rs2076615			
	AA	46.09	42.73	30.86
	AC	50.43	53.64	63.86
	CC	03.48	03.64	05.28
	A	71.30	69.55	62.79
	C	28.70	30.45	37.21
	rs1005753			
	TT	48.72	63.06	46.72
	TG	44.44	35.14	44.10
	GG	06.84	01.80	09.18
	T	70.94	80.63	68.77
	G	29.06	19.37	31.23
PADI4	*rs11203366*			
	GG	21.55	35.96	30.50
	GA	52.59	50.00	47.00
	AA	25.86	14.04	20.90

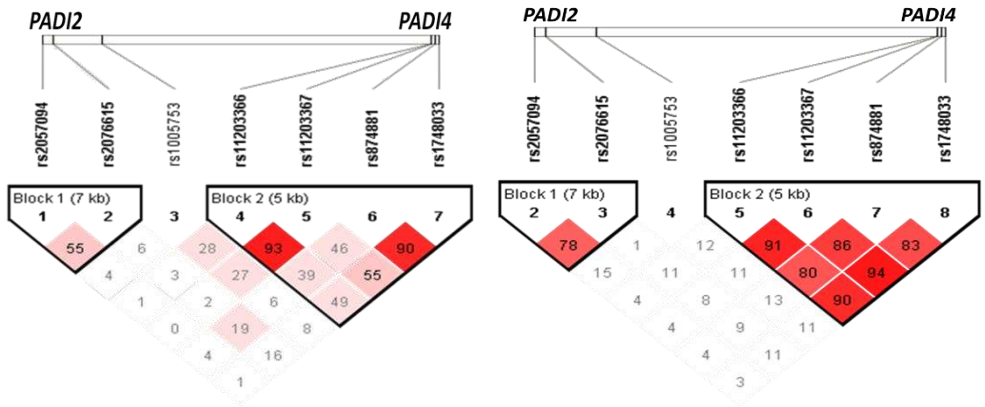
Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 September 2023 doi:10.20944/preprints202309.1802.v1

G	47.84	60.96	54.85
A	52.16	39.04	45.14
rs11203367			
TT	22.52	36.52	29.70
TC	59.46	51.30	48.30
CC	18.02	12.17	20.60
T	52.25	62.17	54.60
C	47.75	37.83	45.40
rs1748033			
CC	16.38	14.91	25.10
CT	50.00	51.57	49.80
TT	33.62	33.33	23.80
C	41.38	40.79	50.66
T	58.62	59.21	49.34
rs874881			
CC	23.28	44.62	24.70
CG	59.48	42.31	51.50
GG	17.24	13.08	22.90
C	53.03	65.77	50.90
G	46.95	34.23	49.10

* F%: Frequency in Percentage.

In Figure 1, diamond plots depict haplotype blocks identified in *PADI2* and *PADI4* within the Mexican mestizo population. The diamonds, connected by lines to form a block, represent a haplotype. At the top of the graph, you can see the 7 SNVs, with each diamond symbolizing the linkage between these loci. Diamonds with a deeper shade of red indicate a higher degree of linkage disequilibrium, while lighter diamonds suggest a lower degree of linkage disequilibrium. The diamonds also display the linkage disequilibrium as 'r' (Pearson's Correlation Coefficient).

The haplotype analysis was conducted between all the groups of patients with RA (with and without ILD) and the group of individuals without the diseases (Figure 1a), as well as within ILD-RA patients vs. RA patients (Figure 1b).



a) b)

Figure 1. Haplotypes in *PADI2* and *PADI4*. a) All Rheumatoid Arthritis patients *vs*. CHS; b) RA-ILD vs. RA patients.

Table 2 displays the percentage frequencies of the haplotypes identified in *PADI4* (rs11203366-rs11203367-rs1748033-rs874881) within the patient populations of ILD-RA, RA and the group of individuals without the diseases (CHS), as shown in Figure 1a,b.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 September 2023 doi:10.20944/preprints202309.1802.v1

Table 2. Analysis of genotypes and alleles of SNV in *PADI4*.

PADI4	*RA-ILD* (*n = 118*, HF%)	*RA* (*n = 133*, HF%)	*CHS* (*n = 616*, *HF%*)
Haplotype	*HF (%)*	*HF (%)*	*HF (%)*
GTTC	39.80	47.40	35.10
ACCG	37.70	32.20	30.80
GTCG	5.00	9.90	13.10
ACTC	7.20	2.90	10.60
GTCC	2.40	1.40	4.00
ATTC	2.30	2.50	1.10
GCCG	2.00	2.40	1.20

HF% = haplotype frequency in percentage.

2.3. AIMs

The genotyping of ancestry informative markers consists of 14 private SNVs between chromosomes 1 and 13. These markers exhibit differential Minor Allele Frequencies (MAF) between the ZAP and CEU populations (taken as reference). Table 3 presents the results of the minor allele frequency in the reference populations, and at the end of the table, you can find the MAF for the study population (n=867 individuals).

Table 3. AIMs polymorphisms used for the calculation of Eigenvectors.

Chr	*SNV*			*ZAP* (*n=60*)			*CEU* (*n=120*)	*D*			*Study group* (*n=867*)
No.	**# rs**	**A1**	**A2**	**MAF**	**A1**	**A2**	**MAF**	**Δ**	**A1**	**A2**	**MAF**
1	rs4528122	T	C	0.067	C	T	0.142	0.792	T	C	0.338
1	rs986690	G	A	0.017	A	G	0.25	0.733	G	A	0.347
4	rs10516422	G	A	0.283	G	A	0.017	0.267	G	A	0.234
5	rs10515716	T	C	0.267	C	T	0.208	0.525	T	C	0.432
6	rs1878071	A	C	0.317	C	A	0.217	0.467	A	C	0.49
9	rs4084051	T	C	0.25	C	T	0.175	0.575	T	C	0.483
9	rs7853112	C	A	0.25	A	C	0.35	0.4	C	A	0.395
9	rs10511491	C	T	0.25	T	C	0.391	0.358	C	T	0.358
9	rs1039336	A	G	0.133	G	A	0.242	0.625	G	A	0.433
9	rs10116714	A	G	0.183	G	A	0.05	0.767	A	G	0.449
9	rs1980888	G	A	0.033	A	G	0.1	0.866	G	A	0.411
9	rs4743556	C	T	0.172	T	C	0.167	0.661	T	C	0.486
12	rs6487927	C	T	0.033	C	T	0.475	0.442	C	T	0.275
13	rs2147155	0	T	0	G	T	0.5	0.5	G	T	0.15

ZAP: Zapotecs, CEU: Utah Residents with Northern and Western European Ancestry, MAF: Minor Allele Frequency, A: allele1 and 2. D: delta value obtained between MAF in CEU and MAF in ZAP populations.

The assessment of genetic distances between the study groups was carried out using the 14 SNVs presented in Table 3. To do this, we conducted a group comparison analysis, and the results of the FST index, along with their corresponding p-values, are shown in Table 4 for the comparisons between the respective row and column.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 September 2023 doi:10.20944/preprints202309.1802.v1

Table 4. Estimated pairwise comparison of Fst index in study groups.

Study group	***RA-ILD***	***RA***
RA	0.32 (0.578)	-
CHS	0.34 (0.853)	0.24 (0.043)

RA: Rheumatoid Arthritis; RA-ILD: RA associated with Interstitial lung disease; CHS: clinically healthy subjects. Show the Fst index (p-value).

2.4. Eigenvectors PCA

Figure 2 displays a Principal Component Analysis (PCA) where the CEU and ZAP populations are taken as reference groups. In this analysis, the control group consists of clinically healthy individuals from the CHS group, and the case group includes all patients from the RA and RA-ILD groups. The first two vectors are shown, collectively providing more than 85% of the information within the population.

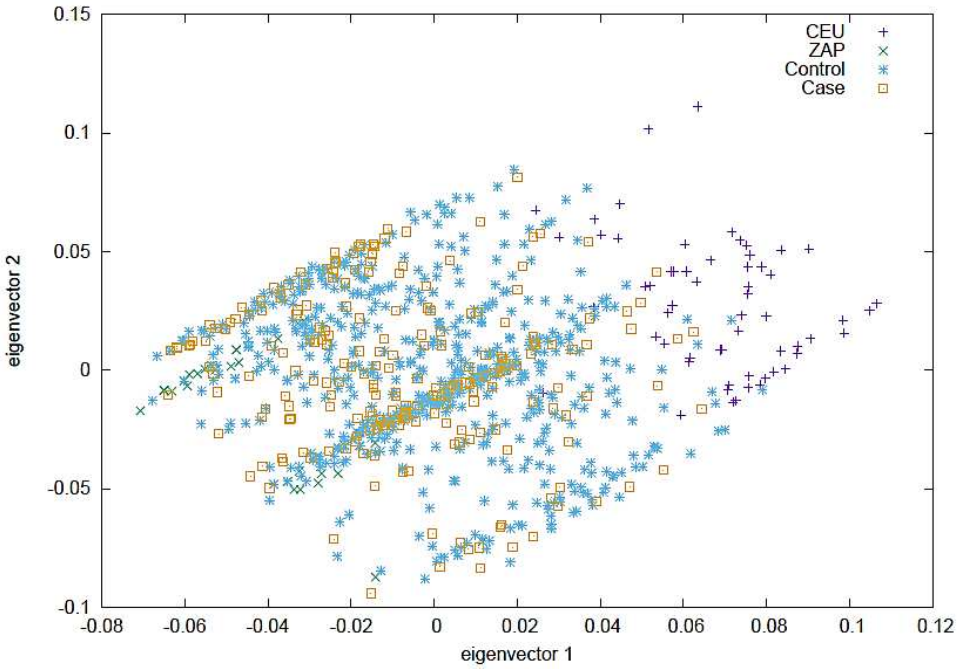
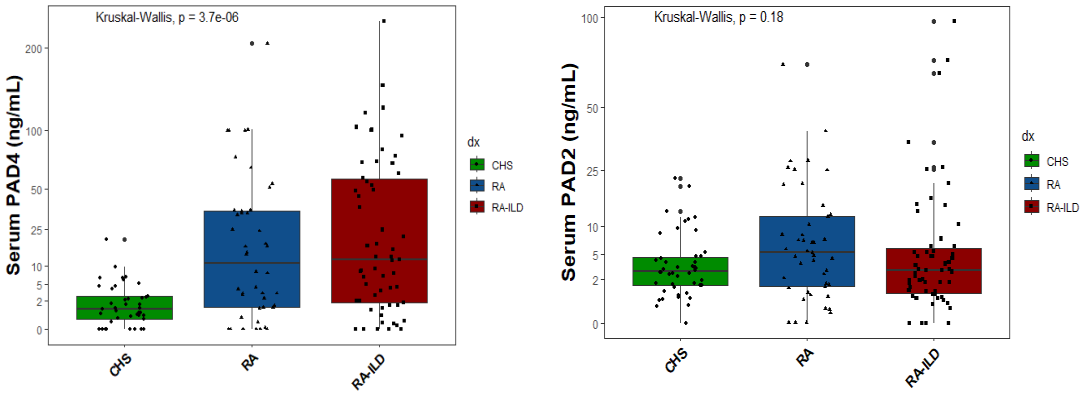


Figure 2. Principal component analysis, populations included in this study; the case group comprises patients with RA-ILD and with only RA (orange square), the control group (blue asterisk), and the reference populations CEU (plus purple symbol) and ZAP (green cross).

2.5. Exposures and protein PAD2 and PAD4 levels

In Figure 3a, the serum levels of PAD4 protein are displayed, and in Figure 3b, the PAD2 protein levels are shown for the three study groups. Each spot (circle, triangle, and black square) represents an individual in whom the protein was measured.

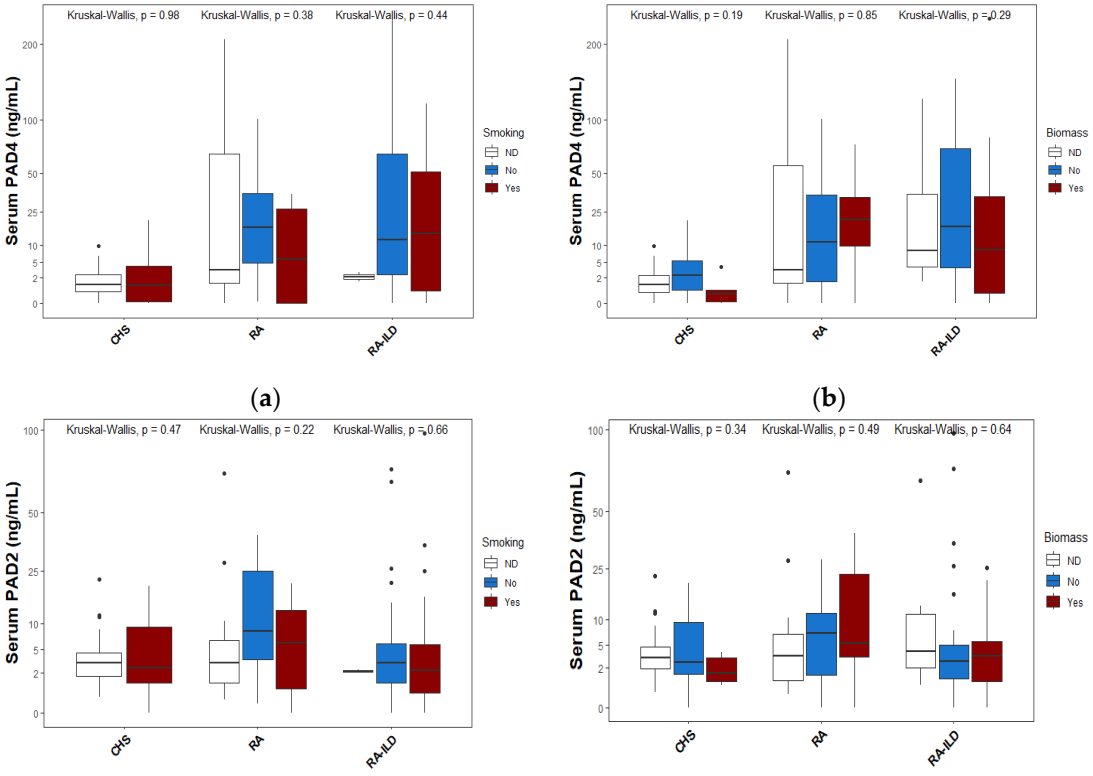
Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 September 2023 doi:10.20944/preprints202309.1802.v1



(**a**) (**b**)

Figure 3. Protein levels determined in serum within the three study groups: a) PAD4 protein levels, b) PAD2 protein levels. Each point, triangle, and square represents an individual. CHS: clinically healthy individuals, RA: rheumatoid arthritis, RA-ILD: rheumatoid arthritis associated with interstitial lung disease.

Figure 4 displays the levels of PAD4 and PAD2 proteins according to exposure to smoking (a and c) or biomass (b and d), which are the primary environmental factors associated with the risk of both lung and joint diseases. In each figure, median values and interquartile ranges are depicted using box-and-whisker plots. The classification is based on smoking status or, in the case of exposure to biomass-burning smoke.



(**a**) (**b**)

(**c**) (**d**)

Figure 4. Protein levels determined in serum within the three study groups, classified by exposure to smoking or biomass: a) PAD4 protein levels in individuals exposed to smoking. b) PAD4 protein levels in individuals exposed to biomass-burning smoke. c) PAD2 protein levels in individuals classified by smoking status. d) PAD2 levels based on biomass status. ND - Data not available.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 September 2023 doi:10.20944/preprints202309.1802.v1

2.6. Proteins in BAL samples

Figure 5 presents the levels of PAD2 and PAD4 proteins, specifically measured in patients diagnosed with RA-ILD who underwent bronchoscopy for diagnostic purposes. Each box in the figure represents an individual patient.

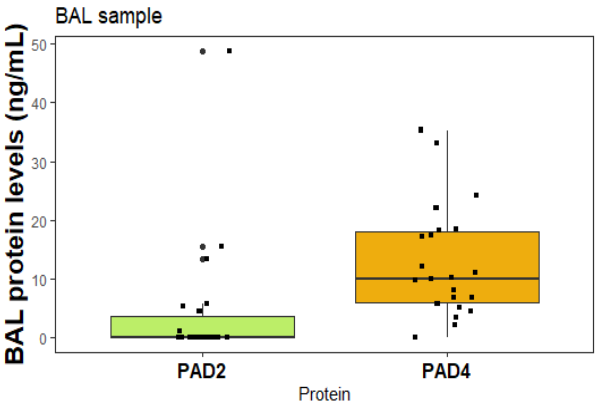


Figure 5. Niveles de proteínas PAD2 y PAD4 determinadas en lavado bronquioalveolar (BAL) en 22 pacientes del grupo RA-ILD.

3. Methods

The genetic data analysis was conducted using Plink v1.07[7]. A total of 23 SNVs were assessed in 907 individuals, with the analysis divided into two parts: initially, 14 AIMs were examined independently, followed by an investigation of 8 SNVs, including four from *PADI2* and four from *PADI4*, for genetic association analysis.

A quality control process was implemented for all polymorphisms to ensure data quality. Individuals with more than 5% missing data, attributed to unsuccessful genotyping of SNVs, were excluded, resulting in the removal of 40 individuals. Additionally, a marker-level evaluation of Hardy-Weinberg equilibrium was performed, leading to the exclusion of SNV rs2235926 in PADI4 ($p = 8.9x10\text{-}24$).

Ancestry markers were analyzed in Plink v1.07, utilizing CEU and ZAP as reference populations, obtained from Phase 3 of the 1000 Genomes Project and the International HapMap Project (websites: http://browser.1000genomes.org and https://ftp.ncbi.nlm.nih.gov/hapmap, respectively) [5,8]. The F_{ST} index was assessed in R using the "Fine Pop" package, and distances were analyzed within a matrix among the three study groups, calculating the global F_{ST} [9].

Protein levels were analyzed through non-parametric tests comparing the three groups. Medians and interquartile ranges were assessed using the Kruskal-Wallis test, and these statistics are presented in box-and-whisker plots for protein levels.

3.1 Ethical statement

Approval for this study was obtained from the Research Institutional Committees for Research, Biosecurity, and Research Ethics at the Instituto Nacional de Enfermedades Respiratorias Ismael Cosío Villegas (INER) under the approval codes B20-15 and C08-15. All participants were granted their authorization and signed informed consent forms. INER provided a document ensuring the safeguarding, including, and adequately handling of personal data as sensitive and confidential information. The research adhered to the guidelines outlined in the 1975 Declaration of Helsinki.

Author Contributions: Conceptualization, Karol Nava-Quiroz, Jorge Rojas-Serrano, Gloria Pérez-Rubio, Ivette Buendia-Roldan and Ramcés Falfán-Valencia; Data curation, Karol Nava-Quiroz, Ivette Buendia-Roldan, Mayra Mejía, Juan Fernández-López and Espiridión Ramos-Martínez; Formal analysis, Karol Nava-Quiroz, Gloria Pérez-Rubio, Mayra Mejía, Juan Fernández-López, Espiridión Ramos-Martínez and Alma Del Ángel-Pablo; Funding acquisition, Ramcés Falfán-Valencia; Investigation, Karol Nava-Quiroz, Jorge Rojas-Serrano, Gloria Pérez-Rubio, Ivette Buendia-Roldan, Mayra Mejía, Juan Fernández-López, Espiridión Ramos-Martínez, Alma

Del Ángel-Pablo and Ramcés Falfán-Valencia; Methodology, Karol Nava-Quiroz, Gloria Pérez-Rubio, Ivette Buendia-Roldan, Juan Fernández-López, Luis López-Flores, Alma Del Ángel-Pablo and Ramcés Falfán-Valencia; Project administration, Ramcés Falfán-Valencia; Resources, Jorge Rojas-Serrano, Mayra Mejía, Espiridión Ramos-Martínez, Luis López-Flores and Ramcés Falfán-Valencia; Software, Karol Nava-Quiroz, Gloria Pérez-Rubio, Juan Fernández-López, Luis López-Flores and Alma Del Ángel-Pablo; Supervision, Jorge Rojas-Serrano, Mayra Mejía and Ramcés Falfán-Valencia; Validation, Jorge Rojas-Serrano, Gloria Pérez-Rubio, Mayra Mejía, Juan Fernández-López, and Alma Del Ángel-Pablo; Visualization, Jorge Rojas-Serrano and Ramcés Falfán-Valencia; Writing – original draft, Karol Nava-Quiroz, Jorge Rojas-Serrano, Gloria Pérez-Rubio, Espiridión Ramos-Martínez and Ramcés Falfán-Valencia; Writing – review & editing, Karol Nava-Quiroz, Gloria Pérez-Rubio and Ramcés Falfán-Valencia.

Institutional Review Board Statement: The research adhered to the principles of the Declaration of Helsinki and received approval from the Ethics Committee of the Instituto Nacional de Enfermedades Respiratorias "Ismael Cosío Villegas" (with protocol codes B20-15 and C08-15).

Informed Consent Statement: Consent was secured from all research participants, encompassing written consent and assurance of personal data confidentiality for publication purposes, both from the patients and healthy subjects who partook in the study.

Data Availability Statement: The datasets generated for this study can be found in ClinVar accessions SCV001422427 – SCV001422434.

Acknowledgments: Programa de Maestría y Doctorado en Ciencias Médicas Odontológicas y de la Salud of the Universidad Nacional Autónoma de México (UNAM). Furthermore, the support provided by the Consejo Nacional de Humanidades, Ciencia y Tecnología (National Council of Science and Technology) (CONAHCyT), CVU: 690362.

Conflicts of Interest: All authors declare no conflict of interest.

References

1. Kelly, C.; Iqbal, K.; Iman-Gutierrez, L.; Evans, P.; Manchegowda, K. Lung Involvement in Inflammatory Rheumatic Diseases. *Best Pract. Res. Clin. Rheumatol.* **2016**, *30*, 870–888, doi:10.1016/j.berh.2016.10.004.
2. Smolen, J.S.; Landewé, R.; Breedveld, F.C.; Buch, M.; Burmester, G.; Dougados, M.; Emery, P.; Gaujoux-Viala, C.; Gossec, L.; Nam, J.; et al. EULAR Recommendations for the Management of Rheumatoid Arthritis with Synthetic and Biological Disease-Modifying Antirheumatic Drugs: 2013 Update. *Ann. Rheum. Dis.* **2014**, *73*, 492–509, doi:10.1136/annrheumdis-2013-204573.
3. Bilgici, A.; Ulusoy, H.; Kuru, O.; Celenk, C.; Ünsal, M.; Danacı, M. Pulmonary Involvement in Rheumatoid Arthritis. *Rheumatol. Int.* **2005**, *25*, 429–435, doi:10.1007/s00296-004-0472-y.
4. Peláez-Ballestas, I.; Sanin, L.H.; Moreno-Montoya, J.; Alvarez-Nemegyei, J.; Burgos-Vargas, R.; Garza-Elizondo, M.; Rodríguez-Amado, J.; Goycochea-Robles, M.V.; Madariaga, M.; Zamudio, J.; et al. Epidemiology of the Rheumatic Diseases in Mexico. A Study of 5 Regions Based on the COPCORD Methodology. *J. Rheumatol.* **2011**, *38*, 3–6, doi:10.3899/jrheum.101024.
5. Silva-Zolezzi, I.; Hidalgo-Miranda, A.; Estrada-Gil, J.; Fernandez-Lopez, J.C.; Uribe-Figueroa, L.; Contreras, A.; Balam-Ortiz, E.; Bosque-Plata, L.; Velazquez-Fernandez, D.; Lara, C.; et al. Analysis of Genomic Diversity in Mexican Mestizo Populations to Develop Genomic Medicine in Mexico. *PNAS* **2009**, *106*, 8611–8616.
6. Regalado, J.; Pérez-Padilla, R.; Sansores, R.; Páramo-Ramirez, J.I.; Brauer, M.; Paré, P.; Vedal, S. The Effect of Biomass Burning on Respiratory Symptoms and Lung Function in Rural Mexican Women. *Am. J. Respir. Crit. Care Med.* **2006**, *174*, 901–905, doi:10.1164/rccm.200503-479OC.
7. Purcell, S.; Neale, B.; Todd-Brown, K.; Thomas, L.; Ferreira, M.; Bender, D.; Maller, J.; Sklar, P.; de Bakker, P.; Daly, MJ & Sham, P. PLINK: A Toolset for Whole-Genome Association and Population-Based Linkage Analysis. *Am. J. Hum. Genet.* **2007**, *81*.
8. Pardo-Seco, J.; Martinón-Torres, F.; Salas, A. Evaluating the Accuracy of AIM Panels at Quantifying Genome Ancestry. *BMC Genomics* **2014**, *15*, doi:10.1186/1471-2164-15-543.
9. Team, R.C. R: A Language and Environment for Statistical Computing. *Vienna, Austria* **2021**.