Novel idiopathic pulmonary fibrosis susceptibility variants revealed by deep sequencing

Jose M. Lorenzo-Salazar[1*], Shwu-Fan Ma[2*], Jonathan Jou[3*], Pei-Chi Hou[2], Beatriz Guillen-Guio[4], Richard J. Allen[5], R. Gisli Jenkins[6], Louise V. Wain[5,7], Justin M. Oldham[8], Imre Noth[2], Carlos Flores[1,4,9]

[1]Genomics Division, Instituto Tecnológico y de Energías Renovables (ITER), Santa Cruz de Tenerife, Spain; [2]Division of Pulmonary and Critical Care Medicine, University of Virginia, Charlottesville, USA; [3]University of Illinois College of Medicine, Chicago, USA; [4]Research Unit, Hospital Universitario N.S. de Candelaria, Universidad de La Laguna, Santa Cruz de Tenerife, Spain;[5]Department of Health Sciences, University of Leicester, Leicester, UK; [6]NIHR Biomedical Research Centre, Respiratory Research Unit, City Campus, University of Nottingham, Nottingham, UK; [7]National Institute for Health Research, Leicester Respiratory Biomedical Research Centre, Glenfield Hospital, Leicester, UK; [8]Pulmonary and Critical Care Medicine, University of California at Davis, Sacramento, USA; [9]CIBER de Enfermedades Respiratorias, Instituto de Salud Carlos III, Madrid, Spain

Supplementary Material

Supplementary Methods

Ethics approval and consent to participate: Discovery study. Institutional review boards and ethics committees at each participating centre approved the study. All participants provided written informed consent. Patients with IPF were clinically characterized at the University of Chicago, from the Correlating Outcomes with biomedical Markers to Estimate Time-progression in IPF (COMET) study, and from and the AntiCoagulant Effectiveness in Idiopathic Pulmonary Fibrosis (ACE-IPF) study. DNA samples were obtained from each individual. The timeframe of sample collection varied by cohort. All DNA samples of individuals with IPF used for association studies were of European-American descent. ACE-IPF and COMET had guidelines for diagnosis of IPF, all of which were adapted from 2000 guidelines from the American Thoracic Society and European Respiratory Society. All patients from the University of Chicago underwent similar diagnostic review in accordance with the 2000 and 2011 guidelines, with each institution engaging in the recommended multidisciplinary (radiology, pathology, and clinical) approach to exclude an alternative diagnosis, as recommended by the 2011 guidelines. All eligible patients were at least 35-years-old and reported having symptoms of idiopathic interstitial pneumonia for at least 3 months. A high-resolution CT scan that showed definite or probable usual interstitial pneumonitis was necessary for inclusion. A surgical lung biopsy sample to confirm usual interstitial pneumonitis could be obtained if the diagnosis was in doubt. Patients with clinically significant exposure to known fibrogenic agents and those with other known causes of interstitial lung disease were excluded before study entry.

Ethics approval and consent to participate: Replication study. The IPF cases and UK Biobank controls used for replication have been previously described [1]. In brief, the IPF cases were

recruited from nine different centres across the UK. IPF diagnoses were made according to the 2002 and 2011 guidelines. Controls from UK Biobank were selected according to the following criteria; they had genome-wide genotype data, were not related to any other individual selected, and presented similar age, sex and smoking distributions to the IPF cases. Additionally, controls with the codes J84.0-J84.9 of the 10th revision of the International Statistical Classification of Diseases and Related Health Problems (ICD10) in the linked Hospital Episodes Statistics data were removed. All individuals were of European ancestry. All the studies were reviewed and approved by the corresponding institution and ethical committee. Informed consent was obtained from all individuals.

Sequencing of the regions-of-interest (ROIs). Genomic DNA (gDNA) was extracted from peripheral blood and the integrity and quantity were assessed. All library preparation reagents were purchased from Agilent Technologies and performed at the Research Resources Centre of University of Illinois at Chicago (Chicago, IL). gDNA was sheared to ~200 bp using Covaris S-220 (Covaris Inc.), and size distribution confirmed by TapeStation analysis. Sequencing (>100X) was performed using the SureSelect™ Target Enrichment System XT2 kit ILM with a custom-designed capture of 1.7 Mb (**Table S1**). Sheared gDNA was then end-repaired, A-tailed and ligated with pre-capture indexing adaptors and PCR amplified. Library sizes and concentration were rechecked. Eight individual indexed libraries were pooled together with equal molar contribution and PCR amplified. The final library pools were quantified by quantitative PCR with the KAPA Biosystems Library Quantification kit (Kapa Biosystems Inc.). Paired-end reads of 100 bases were generated on HiSeq 2500 (Illumina, Inc.).

NGS data processing, variant calling and annotation. SeqPrep 0.4 was used to remove sequence adaptors and merge overlapping reads. NovoAlign 3.02.00 (Novocraft Technologies Sdn

Bhd, Malaysia) was used for read-mapping to the hg19 human genome. Picard 1.7 and SAMtools 0.1.18 [2] were used for BAM file manipulation and to mark duplicates. Qualimap 2.1 [3] indicated that >98% of genomic locations within ROIs were covered at a minimum depth of 90X. GATK 3.3 [4] was used for joint genotyping of small indels and SNVs identified with the HaplotypeCaller following the best practices workflow recommendations. GATK was also used to calculate the transition-to-transversion (Ti/Tv) ratio using the 1000 Genomes Project data (1KGP) as a reference. A false discovery rate (FDR) was calculated to estimate the proportion of false variants declared, assuming a Ti/Tv of 2.25 according to previous whole genome sequencing data [5]. Since the samples were previously genotyped with the Genome-Wide Human SNP Array 6.0 (Affymetrix Inc.) [6], we also evaluated genotyping concordance between array and NGS data. Finally, variant annotation was conducted with ANNOVAR [7] and snpEff4_1h [8] based on data from different sources including the NHLBI Grand Opportunity Exome Sequencing Project, 1KGP, Complete Genomics, COSMIC, dbSNP, and PolyPhen predictions. This information was further supplemented with empirical data generated by the ENCODE project [9] as reported by HaploReg v4.1 [10] and RegulomeDB [11]. Conserved regions that exhibit evidence of selective constraint were also scored by GERP [12] and SiPhy [13].

Validation of low-frequency variants. DNA samples from carriers of the significantly associated rare variants were subjected to direct Sanger sequencing for validation purposes. To overcome the difficulties of the low-complexity sequences within the targeted regions, we used two sets of primer pairs to target the variants in each region: one pair for the amplification step (boost primers) that followed standard PCR protocols, and another pair for sequencing (nested primers), which were used for sequencing both strands of the amplicon to ensure optimal sequencing results. Sequencing was performed by Eurofins Genomics (Louisville, KY) and the resulting traces were

manually revised using 4Peaks v1.8 (Nucleobytes). A list of the primer pairs utilized is provided in **Table S2**.

Replication study. Around 800,000 variants were genotyped in these samples and subjected to strict quality controls. The cases and one third of the controls were genotyped using the Affymetrix Axiom UK BiLEVE array (Affymetrix). The rest of controls were genotyped using the Affymetrix Axiom UK Biobank array (Affymetrix). Phasing and imputation were performed with SHAPEIT v2 [14] and IMPUTE2 v2.3.2 [15], using 1KGP Phase 3 and UK10K as reference panels. Association testing was conducted assuming an additive genetic effect, considering age, sex and the first 10 PCs as covariates. Additional genotyping, imputation and association study details have been described elsewhere [1].

References

1. Allen RJ, Porte J, Braybrooke R, Flores C, Fingerlin TE, Oldham JM, Guillen-Guio B, Ma S-F, Okamoto T, John AE, Obeidat M 'en, Yang IV, Henry A, Hubbard RB, Navaratnam V, Saini G, Thompson N, Booth HL, Hart SP, Hill MR, Hirani N, Maher TM, McAnulty RJ, Millar AB, Molyneaux PL, Parfrey H, Rassl DM, Whyte MKB, Fahy WA, Marshall RP, et al. Genetic variants associated with susceptibility to idiopathic pulmonary fibrosis in people of European ancestry: a genome-wide association study. *Lancet Respir. Med.* 2017; 5: 869–880.

2. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 2009; 25: 2078–2079.

3. Okonechnikov K, Conesa A, García-Alcalde F. Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data. *Bioinformatics* 2016; 32: 292–294.

4. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M, McKenna A, Fennell TJ, Kernytsky AM, Sivachenko AY, Cibulskis K, Gabriel SB, Altshuler D, Daly MJ. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* 2011; 43: 491–498.

5. Li B, Liu DJ, Leal SM. Identifying rare variants associated with complex traits via sequencing. *Curr. Protoc. Hum. Genet.* 2013; Chapter 1: Unit 1.26.

6. Noth I, Zhang Y, Ma S-F, Flores C, Barber M, Huang Y, Broderick SM, Wade MS, Hysi P, Scuirba J, Richards TJ, Juan-Guardela BM, Vij R, Han MK, Martinez FJ, Kossen K, Seiwert SD, Christie JD, Nicolae D, Kaminski N, Garcia JGN. Genetic variants associated with idiopathic pulmonary fibrosis susceptibility and mortality: a genome-wide association study. *Lancet Respir. Med.* 2013; 1: 309–317.

7. Yang H, Wang K. Genomic variant annotation and prioritization with ANNOVAR and wANNOVAR. *Nat. Protoc.* 2015; 10: 1556–1566.

8. Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, Land SJ, Lu X, Ruden DM. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly* 2012; 6: 80–92.

9. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* 2012; 489: 57–74.

10. Ward LD, Kellis M. HaploReg v4: systematic mining of putative causal variants, cell types, regulators and target genes for human complex traits and disease. *Nucleic Acids Res.* 2016; 44: D877–D881.

11. Boyle AP, Hong EL, Hariharan M, Cheng Y, Schaub MA, Kasowski M, Karczewski KJ, Park J, Hitz BC, Weng S, Cherry JM, Snyder M. Annotation of functional variation in personal genomes using RegulomeDB. *Genome Res.* 2012; 22: 1790–1797.

12. Davydov EV, Goode DL, Sirota M, Cooper GM, Sidow A, Batzoglou S. Identifying a high fraction of the human genome to be under selective constraint using GERP++. *PLoS Comput. Biol.* 2010; 6: e1001025.

13. Garber M, Guttman M, Clamp M, Zody MC, Friedman N, Xie X. Identifying novel constrained elements by exploiting biased substitution patterns. *Bioinformatics* 2009; 25: i54–i62.

14. Delaneau O, Marchini J, Zagury JF. A linear complexity phasing method for thousands of genomes. *Nat. Methods* 2011; 9: 179–181.

15. Howie BN, Donnelly P, Marchini J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet.* 2009; 5: e1000529.

Supplementary Tables

Table S1. Three genomic region-of-interest (ROI) and the genes covered by the targeted enrichment design*.

ROI	Size (bp)	Description of targeted region	Genes in the ROI
chr11: 1,213,244-1,430,917[$]	217,674	Centred at *TOLLIP* plus 100 kb upstream and 100 kb downstream	*MUC5AC, MUC5B, TOLLIP, TOLLIP-AS1, BRSK2*
chr14:47,308,828-48,144,457	835,630	Entire *MDGA2* gene	*MDGA2*
chr17:43,672,710-44,327,740	655,031	Centred at *CRHR1, SPPL2C*, and *MAPT* genes plus 25 kb on both ends	*CRHR1, MAPT-AS1, SPPL2C, MAPT, MAPT-IT1, STH, KANSL1*

*Assembled from RefSeq, Ensembl, CCDS, GenCode, and VEGA.
[$]Enrichment excluded the following regions because of design problems due to perfect repeats: chr11:1,307,700-1,309,300 (1600 bp), chr11:1,315,400- 1,316,300 (900 bp), and chr11:1,317,700-1,318,400 (700 bp).

Table S2. Design of amplicons and sequencing primers used to validate rare SNVs using Sanger sequencing.

Targeted SNV(s)	Primer name	Sequence (5' → 3')	Direction	Amplicon size (bp)
rs34474233, r34815853	Boost_Muc5AC_rs4233-5853_F	CTGACCTGCCGACCCAAG	Forward	600
	Boost_Muc5AC_rs4233-5853_R	CTCAGTCCAGAGCCACAGAC	Reverse	
	Nest_Muc5AC_rs4233-5853_F	CTTCCGCAACAGCCTCATC	Forward	301
	Nest_Muc5AC_rs4233-5853_R	CCCCAAAATCCCAGGTGG	Reverse	
rs371630624	Boost_Muc5AC_rs0624_F	CCCCTGTTTCAAAGACCAGC	Forward	543
	Boost_Muc5AC_rs0624_R	AGCAGGTTTGGGTGGAGTAA	Reverse	
	Nest_Muc5AC_rs0624_F	GTGACTGTCATCCTCTGTGC	Forward	197
	Nest_Muc5AC_rs0624_R	ACCCAGGTGTTCAATGTTCAC	Reverse	
rs55938136	Boost_LINC-CRH_rs8136_F	CTGGCTCTTCTCTCTGCTGT	Forward	681
	Boost_LINC-CRH_rs8136_R	CCTGTAATCCCAGCACGTTG	Reverse	
	Nest_LINC-CRH_rs8136_F	TAAGGCCCAATGACACTGTC	Forward	342
	Nest_LINC-CRH_rs8136_R	ATGTAGTGAGACCCTGGCTC	Reverse	
rs543453148	Boost_MDGA2_rs3418_F	CCCTCACTTCTCCTTCCTTTCT	Forward	718
	Boost_MDGA2_rs3418_R	ACAGTTCACGAGGTCAGGAG	Reverse	
	Nest_MDGA2_rs3418_F	CTCATTGCAGCCTCAACTCC	Forward	621
	Nest_MDGA2_rs3418_R	ATCCTGGCTAACACGGTGAA	Reverse	

Table S3. Haplotype results for the top hits of the 11p15.5 region considering risk variants from *MUC5AC*, *MUC5B*, and *TOLLIP* genes. Risk alleles indicated in background grey colour (with and without adjusting for 5 principal components).

	MUC5AC		*MUC5B*		*TOLLIP*				Adjusted		Unadjusted	
Ht#	rs34474233	rs34815853	rs12802931	rs35705950	rs111521887	rs5743894	rs5743890	Freq. (Ca/Co)	OR	P-value	OR	P-value
1	A	A	G	T	C	A	A	0.108/0.028	5.66 (3.20-10.00)	**$2.4x10^{-9}$**	5.08 (2.95-8.74)	**$4.5x10^{-9}$**
2	G	C	G	T	G	G	A	0.159/0.049	3.39 (2.15-5.34)	**$1.3x10^{-7}$**	4.09 (2.64-6.33)	**$2.3x10^{-10}$**
3	G	C	G	T	C	A	A	0.064/0.026	3.66 (1.85-7.22)	**$1.8x10^{-4}$**	3.54 (1.84-6.82)	**$1.6x10^{-4}$**
4	G	C	G	G	C	A	A	0.042/0.026	1.43 (0.67-3.05)	$3.6x10^{-1}$	1.45 (0.70-3.02)	$3.2x10^{-1}$
5	G	C	A	G	C	A	A	0.394/0.576	0.44 (0.33-0.59)	**$5.7x10^{-8}$**	0.40 (0.30-0.53)	**$3.8x10^{-10}$**
6	G	C	A	G	G	G	A	0.079/0.122	0.51 (0.32-0.83)	$7.1x10^{-3}$	0.60 (0.38-0.94)	$2.7x10^{-2}$
7	G	C	G	G	C	A	G	0.021/0.033	0.65 (0.27-1.56)	$3.4x10^{-1}$	0.65 (0.28-1.50)	$3.1x10^{-1}$
8	G	C	A	G	C	A	G	0.094/0.101	0.80 (0.51-1.27)	$3.4x10^{-1}$	0.86 (0.55-1.32)	$4.8x10^{-1}$
9	A	A	G	T				0.131/0.027	6.44 (3.76-11.04)	**$1.3x10^{-11}$**	5.77 (3.45-9.65)	**$2.3x10^{-11}$**
10	G	C	G	T				0.233/0.079	3.70 (2.52-5.43)	**$2.1x10^{-11}$**	4.08 (2.83-5.88)	**$4.8x10^{-14}$**
11	G	C	A	G				0.565/0.805	0.27 (0.20-0.37)	**$1.2x10^{-15}$**	0.26 (0.19-0.35)	**$9.0x10^{-18}$**
12	A	A	A	G				0.000/0.012	0.37 (0.06-2.33)	$2.9x10^{-1}$	0.38 (0.06-2.25)	$2.9x10^{-1}$
13	G	C	G	G				0.064/0.071	0.83 (0.49-1.41)	$5.0x10^{-1}$	0.86 (0.52-1.42)	$5.4x10^{-1}$
14	A	A			C	A	A	0.112/0.042	3.97 (2.35-6.71)	**$2.6x10^{-7}$**	3.62 (2.19-5.99)	**$5.6x10^{-7}$**
15	G	C			G	G	A	0.241/0.185	1.23 (0.91-1.66)	$1.8x10^{-1}$	1.45 (1.08-1.94)	$1.2x10^{-2}$
16	G	C			C	A	A	0.504/0.630	0.62 (0.47-0.82)	**$6.4x10^{-4}$**	0.56 (0.43-0.73)	**$1.4x10^{-5}$**
17	G	C			C	A	G	0.113/0.138	0.76 (0.51-1.14)	$1.8x10^{-1}$	0.80 (0.54-1.17)	$2.4x10^{-1}$
18			G	T	C	A	A	0.176/0.052	6.27 (3.89-10.10)	**$4.3x10^{-14}$**	5.76 (3.65-9.09)	**$5.7x10^{-14}$**
19			G	T	G	G	A	0.179/0.051	4.03 (2.54-6.39)	**$3.0x10^{-9}$**	4.79 (3.08-7.45)	**$3.9x10^{-12}$**
20			G	G	C	A	A	0.043/0.026	1.39 (0.63-3.06)	$4.1x10^{-1}$	1.40 (0.66-2.96)	$3.8x10^{-1}$
21			G	G	G	G	A	0.000/0.015	0.10 (0.01-1.07)	$5.7x10^{-2}$	0.15 (0.02-1.45)	$1.0x10^{-1}$
22			A	G	C	A	A	0.388/0.589	0.42 (0.31-0.57)	**$1.3x10^{-8}$**	0.38 (0.29-0.51)	**$6.8x10^{-11}$**
23			A	G	G	G	A	0.085/0.120	0.53 (0.33-0.86)	$9.8x10^{-3}$	0.62 (0.39-0.97)	$3.6x10^{-2}$
24			G	G	C	A	G	0.021/0.034	0.61 (0.26-1.47)	$2.7x10^{-1}$	0.62 (0.27-1.43)	$2.6x10^{-1}$
25			A	G	C	A	G	0.095/0.103	0.81 (0.51-1.26)	$3.5x10^{-1}$	0.86 (0.56-1.32)	$4.9x10^{-1}$

*Statistical significant combinations in bold.

Supplementary Figures

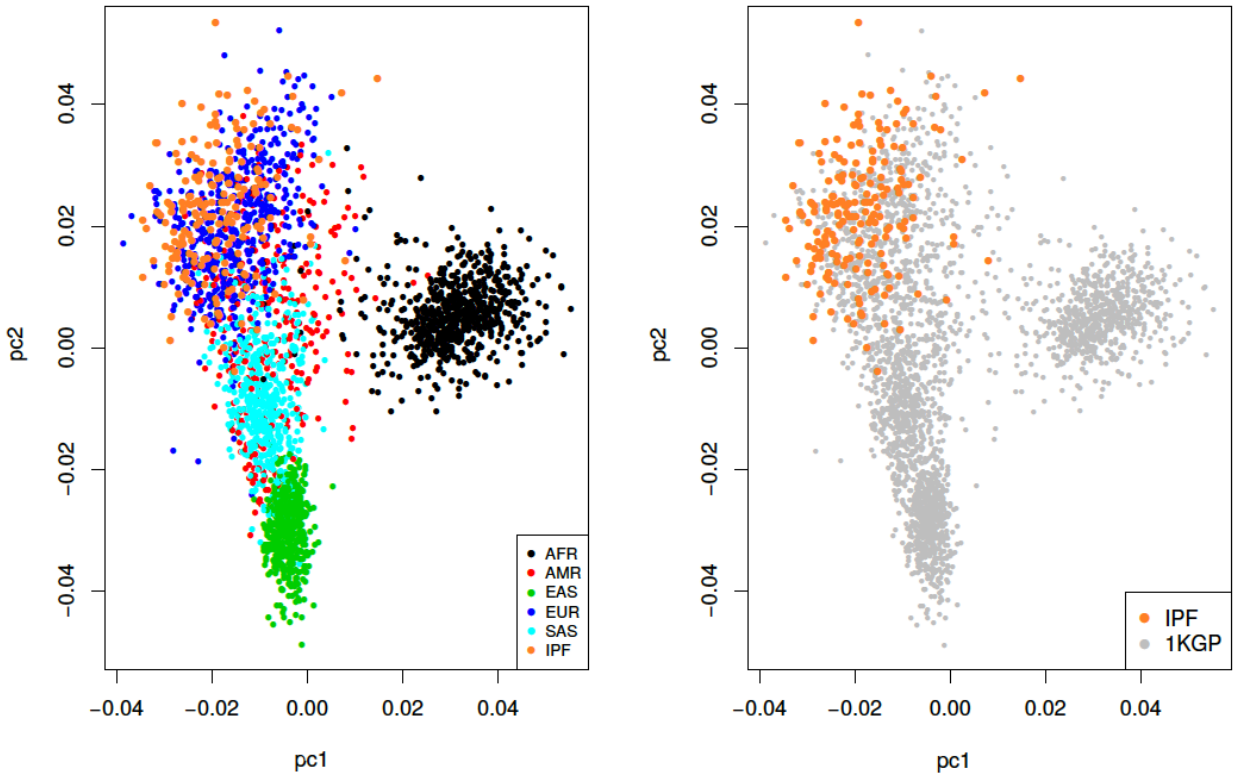

Figure S1. Plot of the first two principal components representing the genetic ancestry scores for IPF cases and the 1000 Genomes Project (1KGP) individuals. IPF cases are indicated with orange circles. AFR, Africans; AMR, admixed Americans; EAS, East Asians; EUR, Europeans; SAS, South Asians.