Genome-wide association study identifies multiple susceptibility loci for pulmonary fibrosis

Tasha E Fingerlin, Elissa Murphy, Weiming Zhang, Anna L Peljto, Kevin K Brown, Mark P Steele, James E Loyd, Gregory P Cosgrove, David Lynch, Steve Groshong, Harold R Collard, Paul J Wolters, Williamson Z Bradford, Karl Kossen, Scott D Seiwert, Roland M du Bois, Christine Kim Garcia, Megan S Devine, Gunnar Gudmundsson, Helgi J Isaksson, Naftali Kaminski, Yingze Zhang, Kevin F Gibson, Lisa H Lancaster, Joy D Cogan, Wendi R Mason, Toby M Maher, Philip L Molyneaux, Athol U Wells, Miriam F Moffatt, Moines Selman, Annie Pardo, Dong Soon Kim, James D Crapo, Barry J Make, Elizabeth A Regan, Dinesha S Walek, Jerry J Daniel, Yoichiro Kamatani, Diana Zelenika, Keith Smith, David McKean, Brent S Pedersen, Janet Talburt, Raven N Kidd, Cheryl R Markin, Kenneth B Beckman, Mark Lathrop, Marvin I Schwarz, David A Schwartz

We performed a genome-wide association study of non-Hispanic, white individuals with fibrotic idiopathic interstitial pneumonias (IIPs; n = 1,616) and controls (n = 4,683), with follow-up replication analyses in 876 cases and 1,890 controls. We confirmed association with TERT at 5p15, MUC5B at 11p15 and the 3q26 region near TERC, and we identified seven newly associated loci (Pmeta = 2.4 × 10−8 to 1.1 × 10−19), including FAM13A (4q22), DSP (6p24), OBF1 (10q24), ATP11A (13q34), DPP9 (19p13) and chromosomal regions 7q22 and 15q14-15. Our results suggest that genes involved in host defense, cell-cell adhesion and DNA repair contribute to risk of fibrotic IIPs.

IIPs represent a group of lung diseases commonly characterized by pulmonary fibrosis or progressive scarring of the alveolar interstitium that can lead to significant morbidity and mortality due to hypoxemic respiratory insufficiency. Although some forms of pulmonary fibrosis are associated with known environmental exposures (for example, asbestos), drug toxicity, radiation exposure or collagen vascular diseases (for example, scleroderma), IIPs have no known etiology. The most common and severe IIP is idiopathic pulmonary fibrosis (IPF), which has a median survival time of 2–3 years after diagnosis. There are no IPF pharmacological therapies approved for use in the United States, and lung transplantation is the only intervention known to prolong life. Although all IIPs have a variable clinical course, they often progress to end-stage lung disease and death. It seems likely that risk of IIP is determined by multiple genetic variants and environmental toxins, but the specific causes of IIP are only beginning to emerge.

The evidence for a genetic component to the risk of IIP is substantial and includes familial clustering of disease, the occurrence of pulmonary fibrosis as part of systemic genetic syndromes, considerable variability in the risk of pulmonary fibrosis among those with similar exposures to known environmental agents, such as asbestos, and the identification of genetic risk loci for IIP. Rare mutations in the TERT, TERC, SFTP C and SFTP A2 genes have been associated with familial interstitial pneumonia (FIP; defined as two or more family members with IIP) and IPF, and a common polymorphism in TERT has been associated with IFP. Recently, we identified a promoter variant in the MUC5B gene (rs35705950) that is present in approximately 50–60% of individuals with FIP or IPF and is estimated to increase risk by 6-fold for heterozygotes and by 20-fold for homozygotes. The identification of MUC5B as a common risk factor has altered our view of the pathogenesis of pulmonary fibrosis from focusing primarily on alveolar epithelia and the lung matrix to inclusion of...
mucus-producing cells in the distal airways of the lung\(^{11,12}\). However, the MUC5B risk variant is observed in \(\sim 19\%\) of unaffected individuals, and approximately one-third of individuals with IIP do not have any identifiable genetic risk factors for this disease, suggesting that other genetic variants contribute to disease risk alone or in combination with the MUC5B variant.

With the goal of identifying additional genetic risk factors that collectively further the understanding of IIPs, we have completed a case-control genome-wide association study (GWAS; 1,616 cases and 4,683 controls) and replication study (876 cases and 1,890 controls) of IIP. We included all types of fibrotic IIP in our case group because (i) distinguishing between the IIP diagnoses is often problematic owing to substantial clinical, pathological and radiological overlap and (ii) there is strong evidence of shared genetic susceptibility, as over 40% of families with FIP have more than one type of IIP among the affected family members\(^3\). We also included both familial and sporadic IIPs because the MUC5B, TERT, TERC and SFTPC variants in individuals with sporadic disease provide suggestive evidence that sporadic IIPs are genetically similar to the familial form of this disease. We hypothesized that susceptibility to IIP is influenced by multiple genetic variants, acting independently or in combination, and that the same genetic variants confer risk of developing different histological types of IIP.

**RESULTS**

**Genome-wide discovery**

We genotyped 1,914 self-reported non-Hispanic, white individuals with fibrotic IIP (cases) on the Illumina Human 660W Quad BeadChip. Of these, 298 were excluded on the basis of being a genetic outlier \((n = 14)\), being a first-degree relative of another case \((n = 126)\), high heterozygosity \((n = 8)\) or missing \(>2\%\) of genotypes across all SNPs \((n = 150)\); Online Methods), such that 1,616 cases were included in the analyses (Supplementary Tables 1–3). Of 15,352 out-of-study controls without phenotypic information also genotyped on the Illumina Human 660W Quad BeadChip in the same laboratory, we selected 4,683 controls who were most genetically similar to our cases on the basis of genome-wide identity-by-state (IBS) comparisons and met the same quality control criteria as cases (Online Methods and Supplementary Table 1).

We compared cases and controls at 439,828 SNPs with (i) minor allele frequency (MAF) of \(>0.05\); (ii) Hardy-Weinberg equi-librium (HWE) \(P\) value of \(>0.0001\) in cases and controls evaluated separately; and (iii) \(P\) value for differential missingness between cases and controls of \(>0.001\) if less than \(2\%\) missing and of \(>0.05\) if between \(2\%\) and \(5\%\) missing. Neither the quantile-quantile plot of \(P\) values (Supplementary Fig. 1) nor the estimated genomic inflation factor \((\lambda)\) of 0.99 suggested any systematic biases, such as those related to population stratification. Under an additive model for the minor allele at each SNP, we identified 19 SNPs, representing 7 chromosomal locations, with genome-wide significant \((P < 5 \times 10^{-8})\) associations (Fig. 1, Table 1 and Supplementary Table 4). In secondary analyses, we identified another genome-wide significant SNP (rs1379326) representing a unique locus under a recessive model (Supplementary Table 4).

**Table 1** Genome-wide significant loci in discovery GWAS and meta-analysis \((P < 5 \times 10^{-8})\)

| SNP       | Locus | Position\(^a\) | Gene\(^b\) | Annotation | Minor allele | MAF case | OR (95% CI) | MAF replication | OR (95% CI) | Meta-analysis | \(P\) |
|-----------|-------|----------------|-----------|------------|--------------|----------|-------------|------------------|-------------|---------------|------|
| rs2736100 | 6p15  | 1339516        | TERT      | Intron     | C            | 0.43     | 0.73 (0.67–0.79) | 7.60 \(\times\) 10\(^{-14}\) | 0.43 | 0.74 (0.65–0.83) | 4.05 \(\times\) 10\(^{-7}\) | 1.71 \(\times\) 10\(^{-19}\) |
| rs2076295 | 6p24  | 7508231        | DSP       | Intron     | G            | 0.54     | 1.43 (1.32–1.55) | 1.14 \(\times\) 10\(^{-16}\) | 0.52 | 1.26 (1.13–1.42) | 6.28 \(\times\) 10\(^{-5}\) | 1.08 \(\times\) 10\(^{-19}\) |
| rs4772443 | 7q22  | 99431282       | TERT      | Intron     | A            | 0.46     | 1.30 (1.20–1.41) | 6.72 \(\times\) 10\(^{-9}\) | 0.42 | 1.11 (0.98–1.24) | 0.093 | 1.17 \(\times\) 10\(^{-8}\) |
| rs7934605 | 11p15 | 1083945        | MUC2      | Intron     | T            | 0.52     | 1.52 (1.40–1.65) | 6.46 \(\times\) 10\(^{-22}\) | 0.51 | 1.56 (1.39–1.76) | 1.49 \(\times\) 10\(^{-13}\) | 6.87 \(\times\) 10\(^{-34}\) |
| rs2034650 | 15q14-15 | 38504594   | TERT      | Intron     | G            | 0.42     | 0.77 (0.71–0.84) | 1.86 \(\times\) 10\(^{-9}\) | 0.42 | 0.82 (0.74–0.93) | 0.00098 | 9.76 \(\times\) 10\(^{-12}\) |
| rs1981997 | 19q12 | 41412603       | MAPT      | Intron     | A            | 0.17     | 0.71 (0.64–0.78) | 2.52 \(\times\) 10\(^{-8}\) | 0.16 | 0.67 (0.58–0.79) | 4.74 \(\times\) 10\(^{-7}\) | 8.87 \(\times\) 10\(^{-14}\) |
| rs12610045 | 19p13 | 46668672       | DPP9      | Intron     | G            | 0.34     | 1.29 (1.18–1.41) | 9.57 \(\times\) 10\(^{-9}\) | 0.34 | 1.30 (1.15–1.47) | 3.94 \(\times\) 10\(^{-5}\) | 1.68 \(\times\) 10\(^{-12}\) |

\(\text{OR}\), odds ratio for the minor allele. The minor allele is defined as the minor allele in the combined case and control group.

\(^a\)Based on NCBI Build 36. \(^b\)Name of gene if SNP falls in body of gene. \(^c\)Adjusted for sex.
Figure 2  Locus-specific plots corresponding to discovery GWAS results for all loci reaching genome-wide significance in the GWAS discovery analysis and meta-analysis of the discovery and replication results. (a–g) For each plot, the $-\log_{10} P$ values (y axis) of the SNPs are shown according to their chromosomal positions (x axis). The significant loci are at 5p15 (a), 6p24 (b), 7q22 (c), 11p15 (d), 15q14-15 (e), 17q21 (f) and 19p13 (g). The estimated recombination rates from the HapMap Project (NCBI Build 36) are shown as blue lines, and the genomic locations of genes within the regions of interest in the NCBI Build 36 human assembly are shown as arrows. SNP color represents LD with the most highly associated SNP at each locus. SNP annotations are as follows: circles, no annotation; squares, synonymous or 3′ UTR; triangles, nonsynonymous; asterisks, TFBScons (in a conserved region predicted to be a transcription factor binding site); squares with an X, MCS44 placental (in a region highly conserved in placental mammals). Genotyped SNPs are shown; analogous plots with imputed SNP genotypes are shown in Supplementary Figure 2.

Replication and meta-analysis

We selected the 20 genome-wide significant SNPs and an additional 178 SNPs with $5 \times 10^{-8} < P < 0.0001$ (143 under an additive model shown between red and blue lines in Fig. 1; see Supplementary Tables 5 and 6 for SNP location, genotype and HWE information and Supplementary Table 4 for association information for all 198 SNPs) for genotyping in a replication cohort of 1,027 cases with IIP and 2,138 controls (Supplementary Tables 1, 3 and 7). After genotype quality control, we included 876 cases and 1,890 controls (Supplementary Tables 1–3) successfully genotyped for 181

Figure 3  Locus-specific plots corresponding to discovery GWAS results for four additional loci reaching genome-wide significance after the meta-analysis of the discovery and replication results. (a–d) For each plot, the $-\log_{10} P$ values (y axis) of the SNPs are shown according to their chromosomal positions (x axis). The significant loci are at 3q26 (a), 4q22 (b), 10q24 (c) and 13q34 (d). The estimated recombination rates from the HapMap Project (NCBI Build 36) are shown as blue lines, and the genomic locations of genes within the regions of interest in the NCBI Build 36 human assembly are shown as arrows. SNP color represents LD with the most highly associated SNP at each locus. SNP annotations are as follows: circles, no annotation; squares, synonymous or 3′ UTR; triangles, nonsynonymous; asterisks, TFBScons; squares with an X, MCS44 placental. Genotyped SNPs are shown; analogous plots with imputed SNP genotypes are shown in Supplementary Figure 3.
of the SNPs. Six of the 8 genome-wide significant loci (13 of 20 SNPs) were associated with IIP in the replication cohort at \( P < 0.0025 \), corresponding to conservative Bonferroni correction for 20 tests (Table 1 and Supplementary Table 4). Seven of the 8 loci (18 of 20 SNPs; Fig. 2) were genome-wide significant in the meta-analysis (Table 1 and Supplementary Table 4). Four additional loci (Fig. 3 and Table 2) were represented among 25 additional SNPs (Supplementary Table 4) that were genome-wide significant under an additive model in the meta-analysis but not in the GWAS discovery.

The most highly associated SNP in the GWAS discovery, rs868903 \((P_{\text{GWAS}} = 1.3 \times 10^{-22}; P_{\text{meta}} = 9.2 \times 10^{-26})\), is in the promoter of the MUC5B gene on chromosome 11p15, which we have previously reported to be associated with IPF and FIP \(^{11} \) and which has been confirmed in other studies \(^{14,15} \). Ten additional SNPs in the MUC5B region, including SNPs in the MUC2 and TOLLIP genes, were also genome-wide significant in the joint analysis and not in strong linkage disequilibrium (LD) with rs868903 (Fig. 2d). The SNPs rs2736100 \( (P_{\text{meta}} = 1.7 \times 10^{-19}) \) and rs2853676 \( (P_{\text{meta}} = 3.3 \times 10^{-8}) \) on chromosome 5p15 are in the TERT gene (Fig. 2a), and rs1881984 \( (P_{\text{meta}} = 4.5 \times 10^{-8}) \) near the TERC gene (Fig. 3a); rare mutations in TERT and TERC have been reported to be associated with FIP and IPF \(^{1,4,5} \), and association of rs2736100 in the TERT gene has previously been reported \(^{10} \).

The remaining eight genome-wide significant loci are newly discovered risk loci for IIP. Five of the association signals, at 4q22, 6p24, 10q24, 13q34 and 19p13, seem localized to single genes (Figs. 2 and 3). SNP rs2609255 \( (P_{\text{meta}} = 2.2 \times 10^{-11}) \) is in the FAM13A gene (family with sequence similarity 13, member A) at 4q22 (Fig. 3b). SNPs rs10484326 \( (P_{\text{meta}} = 5.5 \times 10^{-9}) \) and rs2076295 \( (P_{\text{meta}} = 1.1 \times 10^{-9}) \) are in the DSP gene (desmoplakin) at 6p24 (Fig. 2b). SNPs rs10748858 \( (P_{\text{meta}} = 2.7 \times 10^{-8}) \), rs2067632 \( (P_{\text{meta}} = 3.7 \times 10^{-8}) \) and rs11191865 \( (P_{\text{meta}} = 2.4 \times 10^{-8}) \) are in the OBFC1 gene (oligonucleotide-binding fold containing 1) at 10q24 (Fig. 3c). SNP rs1278769 \( (P_{\text{meta}} = 6.7 \times 10^{-9}) \) is in the ATTP1A gene (ATPase, class VI, type 1A) at 13q34 (Fig. 3d). SNPs rs12610495 \( (P_{\text{meta}} = 1.7 \times 10^{-12}) \) and rs2109069 \( (P_{\text{meta}} = 2.4 \times 10^{-12}) \) are in the DPP9 gene (dipeptidyl-peptidase 9) at 19p13 (Fig. 2g). The other three chromosomal regions (7q22, 15q14-15 and 17q21) have either no significant SNP in any gene or SNPs with significant associations in multiple genes (Fig. 2c,e,f and Tables 1 and 2). The estimated odds ratios (ORs) for all of the genome-wide significant SNPs range from \(-1.1\) to \(-1.6\) (Tables 1 and 2; ORs for SNPs with MAFs that are less than 1 correspond to ORs for the major alleles in this same range).
Imputation across genome-wide significant loci

We imputed genotypes for HapMap 3 SNPs using IMPUTE\textsuperscript{14} across the significantly associated regions to better understand the range over which the association signals extended and to identify additional SNPs potentially associated with IIP (Supplementary Figs. 2 and 3 and Supplementary Table 10). In general, the imputation results were entirely consistent with those from genotyped SNPs. However, the imputation results implicated the *TERC* gene more strongly than the results from direct genotyping (Supplementary Fig. 3a) and seemed to better localize the association signal at 7q22 to the ZKSCAN1 gene (Supplementary Fig. 2c).

Investigation of adjusted models for genome-wide significant SNPs

To adjust for the previously discovered *MUC5B* promoter SNP (rs35705950; not on the Illumina Human 660W Quad BeadChip), we genotyped a subset of the GWAS discovery cases on the same platform and at the same time as the replication cases for the replication SNPs (those listed in Supplementary Table 4). We combined the raw genotypes from these cases (n = 859) with those from the replication cases and controls for joint analyses.

To assess the evidence for multiple independent association signals in each region, we tested for association with each SNP in a given region after adjusting for the most significant SNP in that region on the basis of the meta-analysis results. For the 11p15 region, we adjusted for rs35705950, given our previous findings and the strength of the association we observed between rs35705950 and IIP in our current study population (OR = 4.51, 95% confidence interval (CI) = 3.91–5.21; \( P_{\text{joint}} = 7.21 \times 10^{-95} \)). After adjustment for rs35705950, only one of the SNPs at 11p15 (rs4077759) remained nominally associated with IIP (\( P = 0.03; \) Table 3), whereas rs35705950 remained highly significant in all models (all \( P < 1.81 \times 10^{-89} \)), suggesting that the associations we observed with other SNPs were due to weak LD with rs35705950 (Table 3; see Supplementary Fig. 4 for LD among all the SNPs). The reductions in significance of SNPs in the other regions after adjustment for the top SNP were consistent with LD among the SNPs (Supplementary Table 11) and did not provide evidence for multiple association signals. Notably, SNP rs1881984 near the *TERC* gene was no longer significant after adjustment for rs6793295 in the LRRC34 gene.

Given the increased risk of IIP in males compared to females, we tested for an interaction between each of the GWAS-significant SNPs and sex. We found no strong evidence for differential effects of the SNPs based on sex after correction for the 43 tests (all \( P_{\text{interaction}} > 0.01 \)). Finally, we adjusted for age in addition to sex for all of the genome-wide significant SNPs; with the exception of rs7942850 on chromosome 11 (\( P_{\text{age-adjusted}} = 0.06 \)), all SNPs remained nominally significant after adjustment (Supplementary Table 11).

Expression of key genes in lung tissue

We selected 11 genes for lung tissue expression studies on the basis of localized evidence for newly discovered association with IIP (DPP9, DSP, FAM13A, JVD, DISP2, OBFC1, ATP11A and MUC2) and/or close proximity to an association signal coupled with a priori evidence for expression differences of the gene family in IIP cases compared to controls (*MUC5B, MUC2, WNT3 and WNT9B*).

We measured expression of these genes in lung tissue from 100 IPF cases and 94 controls using quantitative PCR and validated TaqMan Genotyping Assays (Applied Biosystems) to test for differences between cases and controls and to test for association between the genotypes at the most highly associated SNPs in each gene and expression of that gene. We confirmed our results from a smaller study\textsuperscript{11} that *MUC5B* is more highly expressed in the lung tissue of cases compared to controls (\( P = 5.6 \times 10^{-11} \)), but, consistent with our previous findings for rs35705950 in IPF cases, rs868903 was not associated with expression of *MUC5B*. DSP was more highly expressed in cases compared to controls (\( P = 0.0002 \)), and expression differed by genotype at rs2076295 (\( P = 0.002 \)); however, relative expression of DSP decreased with the number of copies of the putative risk allele (Fig. 4). There are two isoforms of desmoplakin generated by alternative splicing. SNP rs2076295 is located in a binding site for transcription factor PU.1, which has been implicated in the alternative splicing of target genes\textsuperscript{19}; however, we saw no evidence for a differential effect of rs2076295 genotype on the expression of the primary isoform compared to the alternative isoform (data not shown). There was nominal evidence for higher expression of DPP9 in cases compared to controls (\( P = 0.03 \)), but neither rs21610495 (\( P = 0.46 \) nor rs2109069 (\( P = 0.72 \)) was associated with DPP9 expression. FAM13A, JVD, OBFC1 and ATP11A did not differ in expression between cases and controls or by genotype (all \( P > 0.12 \)); MUC2, DISP2, WNT3 and WNT9B showed little or no expression in the lung samples.

**Percent variation in disease risk explained by GWAS SNPs**

We estimated the percent of disease risk explained by all 439,828 GWAS SNPs tested for association using a variance components model\textsuperscript{20} across a range of prevalence estimates for IIP (50 per 100,000 to 100 per 100,000). We found that the GWAS SNPs could account for an estimated 28% (standard error (s.e.) of 2%) to 31% (s.e. of 3%) of the risk of IIP. Because we did not include the *MUC5B* promoter SNP (rs35705950) in this analysis (it was not genotyped in the 4,683 individuals in the out-of-study control population), this may be a conservative estimate of the contribution of common SNPs to risk of IIP.

**DISCUSSION**

These findings provide convincing evidence that common genetic variation is an important contributor to risk of IIP. We identified seven new genetic risk loci (4q22, 6p24, 7q22, 10q24, 13q34, 15q14–15 and 19p13) and confirmed the role of risk variants in three previously reported loci (*TERC* at 3q26, *TERT* at 5p15 and *MUC5B* at 11p15).
Before this report, the only common variant consistently associated with IIP was in MUC5B (rs35705950). In aggregate, the common risk variants associated with IIP suggest that this disease is primarily initiated by defects in host defense, cell-cell adhesion and DNA repair. Moreover, our findings can be used to guide intervention trials for this complex disease.

Secreted mucus (MUC5B) in the distal airways seem to have a role in the development of IIP. Our data do not suggest any effects of SNPs in other genes (MUC2 or TOLLIP) in the 11p15 region after accounting for the effect of the MUC5B promoter SNP rs35705950, previously identified as a key risk factor for IIP. SNP rs868903 in the promoter of the MUC5B gene was one of the most strongly associated SNPs in the GWAS, replication and meta-analysis, is not in strong LD ($r^2 = 0.13$) with rs35705950 and is closer to the transcription start site for MUCSB than rs35705950 (1.5 kb versus 3 kb). Although lung tissue from individuals with IIP has higher concentrations of MUC5B than controls, neither of these MUC5B promoter variants seems to be entirely responsible for the increased expression of MUC5B in individuals with IIP; suggesting that other gene variants or environmental toxins are likely to have a role in this disease. We speculate that dysregulated lung mucus initiates or exacerbates lung fibrosis through one of the following mechanisms: (i) altered mucosal defense; (ii) interference with alveolar repair; or (iii) direct cell toxicity (endoplasmic reticulum stress or apoptosis), stimulating a fibroproliferative response initiated by unfolded intracellular MUC5B.

Genes that maintain the length of telomeres seem to have a role in the development of IIP. Before this report, the associations between pulmonary fibrosis and TERT and TERC involved rare variants of TERT and TERC and one common variant of TERT. Mutations in these genes are associated with shortened telomeres in alveolar epithelia, suggesting that these gene variants may increase the risk of pulmonary fibrosis through the disruption of intracellular homeostatic mechanisms. Moreover, dyskeratosis congenita, a congenital disorder that resembles premature aging and frequently involves pulmonary fibrosis, has been attributed to mutations in TERT and TERC. Our GWAS identified common variants in TERT, near TERC and in another gene that influences telomere length, OBFC1. A common variant in OBFC1 has been associated with telomere length in two GWAS of human leukocyte telomere length in the general population. Whether the common variants identified here represent common risk variation or are markers of a collection of rare variants in these genes needs to be established. However, it seems that risk associated with these genes is not limited to rare variants. In aggregate, these findings underscore the importance of telomerase activity, telomere length and possibly early cell senescence in the pathogenesis of pulmonary fibrosis.

Our results implicate alterations in cell-cell adhesion in the risk of developing IIP. Variants in the DSP gene were strongly associated with IIP and with the expression of DSP in the lung tissue of individuals with IIP. DSP encodes desmoplakin, a component of the desmosome, an adhesive intercellular molecule that tightly links adjacent cells and forms a dynamic structure with other proteins (plakoglobin and plakophilins) that tether the cytoskeleton to the cell membrane. Desmosomes are particularly important in maintaining the integrity of tissues that experience mechanical stress (such as the peripheral portions of the lung), and there is strong evidence that perturbation of the desmosome disrupts epithelial homeostasis. Mutations in DSP have been associated with arrhythmogenic right ventricular dysplasia, keratodermas and alopecia, directly implicating desmoplakin in diseases with loss of tissue integrity. More specifically, mutations in DSP have been associated with cardiac interstitial fibrosis on the basis of its overexpression in mouse cardiac tissue. An additional potential mechanism for the involvement of DSP is through alterations in the Wnt/beta-catenin signaling pathway, which have been observed in pulmonary fibrosis. Desmoplakin has been shown to influence the Wnt/beta-catenin signaling pathway through regulation of another component of the desmosome, gamma-catenin. These studies provide strong biomechanical or biological rationales for genetic variation in DSP underlying pulmonary fibrosis.

Our results also implicate other cell-cell adhesion molecules in the risk of IIP development. The DPP9 gene is a member of the same protein family as fibroblast activation protein, which has been shown to be expressed in fibroblastic foci but not in adjacent healthy lung in IIP. DPP9 is expressed in epithelia and has been shown to alter cell adhesion in human embryonic kidney cells. In addition, the CTNNA3 gene (encoding catenin (cadherin-associated protein), alpha 3) was nearly significant in the meta-analysis ($P_{meta} = 9.8 \times 10^{-7}$), is located at 10q22 and is a cell adhesion molecule that physically interacts with beta-catenin and mediates cell adhesion. In aggregate, these findings suggest that pulmonary fibrosis may be caused by defects in cell-cell adhesion or the cytoskeleton that result in an inability to accommodate the stress associated with mechanical stretch of the lung. FAM13A is a signal transduction gene that is responsive to hypoxia, and a SNP (rs7671167) in this gene has recently been found to be protective in chronic obstructive lung disease. The ATP11A gene encodes an ATP-binding cassette (ABC) transporter. Another ABC transporter (ABCA3) is expressed by type II alveolar cells, and mutations in ABCA3 have been shown to interfere with lamellar body formation and surfactant protein function and to cause surfactant protein deficiency in newborns and have been associated with desquamative interstitial pneumonitis and usual interstitial pneumonitis in children and young adults.

The other genome-wide significant loci are not as well localized to a single gene, although there are interesting candidates. At 7q22, an imputed SNP (rs6963345) is the most strongly associated SNP and is in an intron of the ZKSCAN1 gene. ZKSCAN1 is in the same family as ZKSCAN3, variation in which has been associated with airflow obstruction in a large meta-analysis of pulmonary function. The DISP2 gene (encoding dispatched homolog 2) at 15q14-15 encodes a multitransmembrane protein involved in Hedgehog signaling, which is integral to embryogenesis, tissue regeneration and carcinogenesis. The strongest associations at 15q14-15, however, are in or immediately upstream of the JVD gene (encoding isovaleryl-CoA dehydrogenase); JVD is a mitochondrial matrix enzyme involved in leucine catabolism. The association signal at 17q21 was completely confounded by local ancestry in that genomic region, marked by carriage of a common inversion polymorphism, in both the discovery and replication cohorts. As such, determining whether the haplotypes that carry the inversion contain protective variants for IIP will require investigation beyond statistical analysis, such as examination of gene expression differences between cases and controls. An obvious candidate among the genes in the region is WNT7A because alterations in Wnt signaling have been observed in IIP; however, we found no evidence of WNT3 expression in the lung.

Although it has been proposed that pulmonary fibrosis results from activation of developmental pathways or aberrant lung repair, our findings suggest that these mechanisms are secondary to a primary defect in host defense or cell-cell adhesion. Because we found that several genes involved in the integrity of lung epithelia (DSP, DPP9 and CTNNA3) and lung mucus (MUC5B) have IIP risk variants, we hypothesize that defects in these mechanisms primarily contribute to the development of pulmonary fibrosis. Given the importance of...
environmental exposures (for example, exposures to cigarette smoke, asbestos and silica) in the development of interstitial lung disease, it is logical to speculate that common inhaled particles might, over years, cause exaggerated interstitial injury in persons who have defects in lung host defense or cell-cell adhesion. Our view is that shortened telomeres and consequent changes in cell survival in combination with persistent tissue injury may primarily alter host defense or may enhance the ‘host defense challenge’ to the lung through endogenous mechanisms. Thus, excessive lung injury, either through greater environmental exposures, endogenous defects in critical homeostatic mechanisms or subtle defects in host defense, may, over years, lead to pulmonary fibrosis. We believe that more attention should be directed to host defense, cell-cell adhesion and telomere length when considering druggable targets for this complex disease.

Our findings should substantially influence future genetic, diagnostic and pharmacological studies of IIP. We estimated that the cumulative GWAS SNPs (excluding rs35705950) reported in this study explain approximately one-third of the variability in risk of developing IIP, suggesting that further examination of common variation in larger cohorts is warranted, in addition to studies of rare variation, epigenetic features and gene-environment interactions. Although the clinical manifestations of these diseases have been well defined, it is becoming increasingly clear that each type of IIP is influenced by multiple genetic variants that likely have distinct prognoses and may respond differently to pharmacological intervention. Consequently, genotyping subjects with IIP in therapeutic trials may inform drug development by identifying agents that are effective in selected groups of patients. In fact, the lack of attention to pharmacogenetic approaches in IIP trials may explain why few agents have been found to alter the course of these diseases. Moreover, the genetic heterogeneity of IIP suggests that genetic variants may prove helpful in redefining the types of IIP and may provide more accurate prognostic information for afflicted individuals and their families.

URLs. Illumina, http://www.illumina.com; Sequenom, http://www.sequenom.com; SNP3GWA, http://www.phs.wfubmc.edu/public/bios/gene/downloads.cfm; METAL, http://www.sph.umich.edu/csg/abecasis/metal/.

METHODS

Methods and any associated references are available in the online version of the paper.

Note: Supplementary information is available in the online version of the paper.

ACKNOWLEDGMENTS

We gratefully acknowledge the individuals who participated in the studies that contributed to this work. This research was supported by the National Heart, Lung, and Blood Institute (R01-HL095393, R01-HL097163, P01-HL092870, R2C-HL101715, U01-HL089897, U01-HL089856, U01-HL108642 and P50-HL089432), the Veterans Administration (I10BX001534), the Dorothy F and Richard P Simmons Center and InterMune.

AUTHOR CONTRIBUTIONS

T.E.F. and D.A.S. designed the study. K.K.B., M.P.S., J.E.L., G.P.C., D.L.G., S.G., H.R.C., K.K., P.W.j, R.M.d.B., C.K.G., M.S.D., G.G., H.J.I., N.K., Y.Z., K.F.G., L.H.L., T.E.F., W.Z., A.L.P., B.S.P. and Y.K. analyzed the data. T.E.F, M.L. and D.A.S. developed the conceptual approaches to data analysis. T.E.F. and D.A.S. wrote the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare competing financial interests: details are available in the online version of the paper.

Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. American Thoracic Society & European Respiratory Society. American Thoracic Society/European Respiratory Society International Multidisciplinary Consensus Classification of the Idiopathic Interstitial Pneumonias. Am. J. Respir. Crit. Care Med. 165, 277–304 (2002).
2. Hubbard, R. et al. Exposure to commonly prescribed drugs and the etiology of cryptogenic fibrosing alveolitis: a case-control study. Am. J. Respir. Crit. Care Med. 157, 743–747 (1998).
3. Raghu, G. Idiopathic pulmonary fibrosis: new evidence and an improved standard of care in 2012. Lancet 380, 699–701 (2012).
4. Armanios, M.Y. et al. Telomerase mutations in families with idiopathic pulmonary fibrosis. N. Engl. J. Med. 355, 2372–2380 (2006).
5. Tsakiri, K.D. et al. Adult-onset pulmonary fibrosis caused by mutations in telomerase. Proc. Natl. Acad. Sci. USA 104, 7552–7557 (2007).
6. Thomas, A.Q. et al. Heterozygosity for a surfactant protein C gene mutation associated with usual interstitial pneumonitis and cellular nonspecific interstitial pneumonitis in one kindred. Am. J. Respir. Crit. Care Med. 165, 1322–1328 (2002).
7. Lawson, W.E. et al. Genetic mutations in surfactant protein C are a rare cause of sporadic cases of IPF. Thorax 59, 977–980 (2004).
8. Wang, Y. et al. Genetic defects in surfactant protein A2 are associated with pulmonary fibrosis and lung cancer. Am. J. Hum. Genet. 84, 52–59 (2009).
9. van Noostrum, C.H. et al. Surfactant protein C mutations are the basis of a significant portion of adult familial pulmonary fibrosis in a Dutch cohort. Am. J. Respir. Crit. Care Med. 182, 1419–1425 (2010).
10. Mushiroda, T. et al. A genome-wide association study identifies an association of a common variant in TERT with susceptibility to idiopathic pulmonary fibrosis. J. Med. Genet. 45, 654–656 (2008).
11. Seibold, M.A. et al. A common MUC5B promoter polymorphism and pulmonary fibrosis. N. Engl. J. Med. 364, 1503–1512 (2011).
12. Boucher, R.C. Idiopathic pulmonary fibrosis—a sticky business. N. Engl. J. Med. 364, 1560–1561 (2011).
13. Steele, M.P. et al. Clinical and pathologic features of familial interstitial pneumonia. Am. J. Respir. Crit. Care Med. 172, 1146–1152 (2005).
14. Zhang, Y., Noth, I., Garcia, J.G. & Kaminski, N. A variant in the promoter of MUC5B and idiopathic pulmonary fibrosis. N. Engl. J. Med. 364, 1576–1577 (2011).
15. Stock, C.J. et al. Mucin 5B promoter polymorphism is associated with idiopathic pulmonary fibrosis but not with development of lung fibrosis in systemic sclerosis or scleroderma. Thorax published online; doi:10.1136/thoraxjnl-2012-207186 (15 January 2013).
16. Boettger, L.M., Handsaker, R.E., Zody, M.C. & Carroll, S.A. Structural haplotypes and recent evolution of the human 17q21.31 region. Nat. Genet. 44, 881–885 (2012).
17. Steinberg, K.M. et al. Structural diversity and African origin of the 17q21.31 inversion polymorphism. Nat. Genet. 44, 872–880 (2012).
18. Marchini, J., Howie, B., Myers, S., McVean, G. & Donnelly, P. A new multpoint method for genome-wide association studies by imputation of genotypes. Nat. Genet. 39, 906–913 (2007).
19. Guilford, C., Gallais, I. & Moreau-Gachelin, F. Spi-1/PU.1 oncogene affects splicing decisions in a promoter binding-dependent manner. J. Biol. Chem. 281, 19145–19155 (2006).
20. Yang, J., Lee, S.H., Goddard, M.E. & Visscher, P.M. GCTA: a tool for genome-wide complex trait analysis. Am. J. Hum. Genet. 86, 11–21 (2010).
21. Rose, M.C. & Voynow, J.A. Respiratory tract mucin genes and mucin glycoproteins in health and disease. Physiol. Rev. 86, 245–278 (2006).
22. Crouch, E. Pathobiology of pulmonary fibrosis. Am. J. Physiol. (Lung Cell. Mol. Physiol.) 259, L159–L184 (1990).
23. Kuhn, C. et al. An immunohistochemical study of architectural remodeling and connective tissue synthesis in pulmonary fibrosis. Am. Rev. Respir. Dis. 140, 1693–1703 (1989).
24. Korfel, M. et al. Epithelial endoplasmic reticulum stress and apoptosis in sporadic idiopathic pulmonary fibrosis. Am. J. Respir. Crit. Care Med. 178, 838–846 (2008).
25. Alder, J.K. et al. Short telomeres are a risk factor for idiopathic pulmonary fibrosis. Proc. Natl. Acad. Sci. USA 105, 13051–13056 (2008).
26. Levy, D. et al. Genome-wide association identifies ZFCH1 as a locus involved in human leukocyte telomere biology. Proc. Natl. Acad. Sci. USA 107, 9293–9298 (2010).
27. Meirhofer, M. et al. Genome-wide meta-analysis points to CTCL and ZNF676 as genes regulating teleome homeostasis in humans. Hum. Mol. Genet. 21, 5385–5394 (2012).
Compound heterozygosity for non-sense and mis-sense variants in Desmoplakin acts as a tumor suppressor by inhibition of the Wnt/β-catenin signaling pathway in human lung cancer.
ONLINE METHODS

Study populations. We used standard criteria established by the American Thoracic Society/European Respiratory Society\(^1\) to determine the diagnostic classification of all cases in the discovery and replication phases (Supplementary Tables 1–3 and 7). We excluded cases with known explanations for the development of fibrotic IIP, including infections, systemic disorders or relevant exposures (for example, asbestos). To maximize power and minimize potential confounding by ancestry, we included only self-reported non-Hispanic, white participants in the GWAS and replication studies. All subjects gave written informed consent as part of institutional review board (IRB)-approved protocols for their recruitment, and the GWAS was approved by the National Jewish Health IRB and the Colorado Combined Institutional Review Board (COMIRB).

GWAS discovery. We genotyped 1,914 subjects with IIP from 6 cohorts (FIP (n = 566), National Jewish Health IIP population (n = 238), InterMune IPF trials (n = 720), UCSF (n = 66), Vanderbilt University IIP population (n = 105) and the National Heart, Lung, and Blood Institute Lung Tissue Research Consortium (n = 219)) and compared them to genotypes from 4,683 out-of-study controls (Supplementary Tables 1–3). After genotype quality control, we included 1,616 cases in analyses.

A family with FIP was defined by the presence of at least two cases of definite or probable IIP in individuals who are third-degree relatives or closer. Recruitment of families based at three major referral centers (Vanderbilt University, Duke University and National Jewish Health) has been ongoing since 1999. We included only one IIP case among first-degree relatives. The National Jewish Health IIP cohort consists of individuals with sporadic IIP who were clinically evaluated and enrolled at National Jewish Health as part of ongoing research protocols associated with clinical care. Details of the recruitment criteria for the cases from the InterMune IPF \(\gamma\)-Interferon Intervention Trial have been described in detail\(^5\). Briefly, eligible subjects had IPF and were 40 to 79 years old with clinical symptoms for at least 3 months and evidence of disease progression within the previous 12 months. We included all available cases, regardless of treatment assignment. The National Heart, Lung, and Blood Institute Lung Tissue Research Consortium (NHLBI LTRC) was established to provide lung tissue and DNA for the research community. We included DNA from those subjects with a diagnosis of IIP.

We used deidentified control genotypes generated at Centre d’Etude du Polymorphisme Humain (CEPH) as part of other studies. Potential controls were those who self-reported non-Hispanic, white ancestry, had been genotyped at the Centre national de Séquençage des Populations (CEPH) using the Illumina Human 660W Quad BeadChip. Barcoded DNA samples were received in standard tubes together with sample information and were subjected to stringent quality control. Concentration, fragmentation and response to PCR were determined. Samples from cases and controls were randomly distributed on 96-well plates. Processing was carried out under full LIMS (laboratory information management system) control in a fully automated Illumina BeadLab equipped with eight Tecan liquid-handling robots, six Illumina BeadArray readers and two Illumina iScans.

Replication. Replication genotyping was carried out at CEPH using the Illumina Human 660W Quad BeadChip. Barcoded DNA samples were received in standard tubes together with sample information and were subjected to stringent quality control. Concentration, fragmentation and response to PCR were determined. Samples from cases and controls were randomly distributed on 96-well plates. Processing was carried out under full LIMS (laboratory information management system) control in a fully automated Illumina BeadLab equipped with eight Tecan liquid-handling robots, six Illumina BeadArray readers and two Illumina iScans.

Validation genotyping was accomplished with a combination of multiplexed (Sequenom iPLEX) and uniplex (TaqMan) assays. First, assay design for multiplexed Sequenom iPLEX genotyping was performed for an input set of 198 SNPs (Supplementary Table 4), using a combination of web-based (AssayDesigner Suite) and desktop (AssayDesigner) software tools (Sequenom). Of 198 input SNPs, 193 were efficiently placed into a set of 6 assays of the following plexities: 35, 35, 35, 35, 31 and 22 SNPs. Sequenom iPLEX genotyping is based on multiplexed locus-specific PCR amplification, multiplexed single-base extension (SBE) from locus-specific amplicons and multiplexed resolution of SBE product base calling using matrix-assisted laser desorption ionization–time of flight (MALDI-TOF) mass spectrometry.

Primers for the Sequenom assays were purchased from IDT, and all steps of the iPLEX procedure were carried out using reagents and methods from Sequenom according to the manufacturer’s instructions. Reactions were carried out in 384-well plates and analyzed using the Sequenom MassARRAY Analyzer 4 system with iPLEX Gold reagents and SpectroCHIP arrays. Results were analyzed using a combination of commercial software (Typer 4, Sequenom) and custom tools for data management. Of 193 assays in 6 multiplexes, 176 were successful in generating usable genotyping data.

DNA preparation, storage and quality control. Genomic DNA was isolated from whole blood and biopsied lung tissue on either the Autopure LS (Qiagen) or QIAcube (Qiagen) automation platform, respectively. Before extraction on the QIACube using the DNeasy kit, fibrotic lung tissues were first homogenized using Lysing Matrix D tubes and a FastPrep-24 benchtop homogenizer (MPBiomedicals). After isolation, all DNA was assayed for concentration and purity on a NanoDrop ND-1000 Spectrophotometer (Thermo Scientific). Samples were excluded if DNA had a concentration of <50 ng/\(\mu l\) or an \(A_{260}/A_{280}\) ratio outside the 1.7–2.0 range.

For GWAS genotyping, before submission to the Centre National de Genotypage (CNG) at CEPH, all samples were requantified using the Quant-IT PicoGreen dsDNA Assay kit (Invitrogen), normalized with 1× TE and aliquoted into individually barcoded screw-cap tubes. Owing to volume limitations with liquid-handling robots, an absolute minimum quantity for submission to CNG was 30 µl at a concentration of 50 ng/µl. If samples did not meet this minimum quantity, an alternate extraction was performed or the sample was withheld from the study.

For replication genotyping, upon receipt, samples were transferred into 96-well robotics-compatible plates, quantified with PicoGreen and normalized with 1× TE. DNA (400 ng) was submitted for each subject in the GWAS, and the replication cohorts were sent for replication genotyping. In an effort to minimize confounding by batch effects, samples were aliquotted into 96-well plates in a randomized fashion across all cohorts with two duplicates per plate using the Tecan Evo200 liquid-handling robot.

Genome-wide genotyping. Genome-wide genotyping was carried out at CEPH using the Illumina Human 660W Quad BeadChip. Barcoded DNA samples were received in standard tubes together with sample information and were subjected to stringent quality control. Concentration, fragmentation and response to PCR were determined. Samples from cases and controls were randomly distributed on 96-well plates. Processing was carried out under full LIMS (laboratory information management system) control in a fully automated Illumina BeadLab equipped with eight Tecan liquid-handling robots, six Illumina BeadArray readers and two Illumina iScans.

Replication genotyping. Replication genotyping was carried out at the Biomedical Genomics Center at the University of Minnesota. We genotyped 198 SNPs with \(P\) values less than 0.0001 (see Statistical Analyses) in 1,027 independent cases and 2,138 controls. We also genotyped the MUC5B promoter SNP rs35705950, which is not on the Illumina Human 660W Quad BeadChip, to allow adjustment of other SNPs at \(P \leq 5\) for rs35705950. In addition, to allow follow-up joint statistical tests (using raw genotypes from both GWAS cases and replication cases and controls) with adjustment for covariates that were not available for the out-of-study controls, we also genotyped a subset of GWAS cases. Details of the validation assays are described below. After genotyping quality control, we included 876 cases and 1,890 controls in the replication, meta- and joint analyses and 859 of the GWAS cases in the joint analyses.

Before genotyping, quality control was performed on all samples by quantitative RT-PCR and uniplex genotyping using TaqMan. Samples that did not pass either quality control step, although carried forward through genotyping, were later removed from analysis.

Before genotyping, quality control was performed on all samples by quantitative RT-PCR and uniplex genotyping using TaqMan. Samples that did not pass either quality control step, although carried forward through genotyping, were later removed from analysis.
The remaining five SNPs that were not successfully imputed in the original Sequenom iPLEX designs (rs2736100, rs13225346, rs10822856, rs10193981 and rs10751635), as well as a sixth SNP (rs35705950) published in earlier studies, were genotyped using commercial TaqMan assays (Life Technologies). Reactions were carried out in 384-well plates, and fluorescence was read using an Applied Biosystems ABI 7900HT Sequence Detection System.

**Gene expression.** Total RNA was isolated from approximately 30 mg of snap-frozen or RNAlater-preserved lung tissue using the Ambion mirVana kit (Life Technologies). RNA concentration was determined by NanoDrop ND-1000, and RNA integrity was determined using the 2100 BioAnalyzer (Agilent). cDNA single-strand conversion was performed using the Superscript III First-Strand Synthesis System (Invitrogen), and expression analysis was carried out using predesigned TaqMan assays run on the Viia7 Real-Time PCR instrument (Life Technologies): DPP9, Hs00373589; DSP, Hs00189422; DSP variant 1, Hs00950584; FAM13A, Hs00208453; IJD, Hs01064832; MUC5B, Hs00861588; MUC2, Hs00149374; OBFC1, Hs00998858; WNT3, Hs00902257; WNT9B, Hs00166462; GAPDH, 4333764F; ATP11A, Hs00392589; DISP2, Hs01592340_m1. All assays were run in triplicate with GAPDH used as the endogenous control. As an additional control, one sample per plate was run in duplicate from the cDNA conversion step.

**Statistical analyses.** Selection of out-of-study controls for GWAS discovery. We obtained controls on the basis of genetic matching to cases from a large database of anonymous genotypes from Europeans who had been genotyped at CEPH on the Illumina Human 660W Quad BeadChip. An ancestry analysis was carried out using EIGENSTRAT3.0 software. HapMap data on 618 individuals (CEU, YRI, JPT and CHB) and samples of reference Europeans were used as representatives of European, West African and East Asian populations to infer ancestry-informative principal components that were projected onto the cases and control samples. Putative non-European samples were flagged as outliers and removed from all subsequent analyses. We selected a subset of the available controls to obtain three matching controls per case by first jointly clustering the cases and controls into subpopulations on the basis of the top ten principal components using a support vector machine approach (R package e1071 with radial basis function). We then applied an optimal paired matching algorithm within each cluster to choose the best three controls for each case on the basis of a distance matrix defined by the top ten principal components (pairmatch function in the R package optmatch). Removal of first-degree relatives in cases and controls. We included only individuals from first-degree relatives identified on the basis of an estimated kinship coefficient of ≥0.05. For estimation of the percent variation in disease risk explained by the GWAS SNPs, which is particularly sensitive to cryptic relatedness, we kept only one individual of those with estimated kinship coefficient of >0.025. Exclusion of individuals and prioritization of SNPs for the discovery GWAS. In addition to individuals excluded by the CNG at CEPH, we excluded cases and controls with (i) evidence of being a genetic outlier based on a pairwise IBS estimate with the 5th closest neighbor that was >4 s.d. from the mean pairwise IBS estimate across all pairs; (ii) unresolved sex mismatch between clinical and genomic data; (iii) heterozygosity across the SNPs greater or less than 4 s.d. from the mean heterozygosity based on the distribution of heterozygosity across all individuals; and (iv) genotype calls at less than 98% of SNPs that passed laboratory quality control. On the basis of this quality control, we excluded 298 cases and 145 controls. In addition to the laboratory quality control measures, we prioritized SNPs for follow-up on the basis of other criteria. We tested for differential missingness via a χ² test of proportions of missingness between cases and controls and for departures from HWE via a 1-degree-of-freedom goodness-of-fit test. We prioritized SNPs with (i) MAF of >0.05; (ii) HWE P of >0.0001 in cases and controls evaluated separately; and (iii) P value for differential missingness between cases and controls of >0.001 if less than 2% missing and >0.05 if between 2% and 5% missing.

GWAS association testing. We tested for association between each SNP and IIP using an exact mixed-model approach to account for subtle relatedness and population stratification among our cases and controls implemented in the genome-wide efficient mixed-model association (GEMMA) software package. We tested for association under an additive model for our primary analysis and, in a secondary analysis, took the minimum of the recessive and dominant models P values if there was significant lack of fit to the additive model (P < 0.05) from a linear regression that assumed independence among the samples (such a test is not currently implementable in the GEMMA software). We adjusted for sex in all models. We compared the distribution of P values obtained under the additive model to that expected under the null hypothesis of no association across the genome and report the quantile-quantile plot and genomic inflation factor (λ) to verify the absence of systematic biases due to experimental or other confounding factors such as population stratification. We selected all SNPs with P value < 0.0001 for follow-up in the replication populations. We visually inspected genotype spectra for all 198 selected SNPs to assure genotype call quality. We calculated ORs and 95% CIs from a logistic regression model adjusted for sex, assuming independence among the cases and controls, as the linear model in GEMMA uses the identity link rather than the log-odds link function. As such, the CIs may be slightly narrower than those based on full mixed models.

**Replication association.** We tested for association between each replication SNP and IIP in the replication cases and controls using freely available SNPGWA software (see URLs). We tested for association under the genetic model from the GWAS that gave the minimum P value (143 under an additive model, 24 under a dominant model and 31 under a recessive model). P < 0.0025 was considered to represent statistically significant replication for the 20 genome-wide significant GWAS SNPs. The P values for the other 178 SNPs were used in the meta-analysis of the GWAS and replication cohorts.

**Meta-analysis.** To obtain an overall measure of association between each of the 181 successfully genotyped SNPs in the replication set and IIP, we performed a meta-analysis of the GWAS and replication results. We used the weighted inverse normal method, letting Zᵢ (i = GWAS or replication) be the test statistic from the test of association in the ith study and letting vᵢ (i = GWAS or replication) be the corresponding weight. We took the weight to be the square root of the total sample size in the ith study because effect estimates from the GWAS and replication phases were not on the same scale. Note that this method explicitly accounts for the directionality of the association. Thus, highly significant associations with conflicting directions do not exhibit strong statistical association in this meta-analysis. We used METAL to perform our meta-analysis. SNPs with Pmeta < 5 × 10⁻⁸ were considered genome-wide significant. We created locus-specific plots of the discovery GWAS results for all loci that were genome-wide significant in the meta-analysis.

**Stratified analyses by H1 and H2 haplotypes on chromosome 17.** We stratified our discovery and replication cohorts using rs17563986, which completely determines the H1 or H2 haplotype. We tested for association with individuals carrying two H1 haplotypes, given the small sample sizes of H2 carriers. There were 1,127 cases and 2,832 controls included in the stratified GWAS analysis and 617 cases and 1,138 controls in the stratified replication analysis.

**Imputation.** We imputed genotypes using the combined case and control discovery samples for all HapMap SNPs across a 5-Mb region that covered the association signal for each genome-wide significant locus. We used multi-population reference panel data from HapMap 3 for phasing using Shapeit with appropriate default parameters. We performed imputation using IMPUTE and tested for association at only those SNPs with imputation information as measured by .info of >0.5 using SNFTEST (v2) with multiple Newton-Raphson iterations to estimate parameters.

**Multi-SNP models.** To assess the independence of effects of the meta-analysis genome-wide significant SNPs, we used logistic regression models within each locus using a combined case group (subset of GWAS and all replication) and the replication controls using SAS (v. 9.2). Specifically, within each locus with a genome-wide significant SNP, we tested for association between IIP and each of the other validation panel SNPs within that locus after adjusting
for the most significantly associated SNP in that locus (at 11p15, we adjusted for rs35705950). To assess the robustness of each SNP association to age effects in addition to sex, we tested for association between IIP and each SNP adjusted for both age and sex. Finally, to test for effect modification of SNP associations by sex, we tested for association between IIP and each SNP by sex interaction.

Expression analyses. We tested for differential gene expression in the lung between 100 cases and 94 controls using a 2-sample t test. We also tested for differential expression by genotype using the combined case and control group via a test for trend across the three genotype groups unless there were <5 individuals in a genotype group; we grouped the rare homozygote and heterozygote groups together in that case. \( P < 0.05 \) was considered statistically significant.

53. King, T.E. Jr. et al. Effect of interferon \( \gamma \)-1b on survival in patients with idiopathic pulmonary fibrosis (INSPIRE): a multicentre, randomised, placebo-controlled trial. Lancet 374, 222–228 (2009).
54. Regan, E.A. et al. Genetic epidemiology of COPD (COPDGene) study design. COPD 7, 32–43 (2010).
55. Patterson, N., Price, A.L. & Reich, D. Population structure and eigenanalysis. PLoS Genet. 2, e190 (2006).
56. Heath, S.C. et al. Investigation of the fine structure of European populations with applications to disease association studies. Eur. J. Hum. Genet. 16, 1413–1429 (2008).
57. Cortes, C. & Vapnik, V. Support-vector networks. Mach. Learn. 20, 273–297 (1995).
58. Hansen, B.B. & Klopfer, S.O. Optimal full matching and related designs via network flows. J. Comput. Graph. Statist. 15, 1–19 (2006).
59. Harley, J.B. et al. Genome-wide association scan in women with systemic lupus erythematosus identifies susceptibility variants in \( iT G A M \), \( P X K \), \( K I A A 1 5 4 2 \) and other loci. Nat. Genet. 40, 204–210 (2008).
60. Zhou, X. & Stephens, M. Genome-wide efficient mixed-model analysis for association studies. Nat. Genet. 44, 821–824 (2012).
61. Willer, C.J., Li, Y. & Abecasis, G.R. METAL: fast and efficient meta-analysis of genomewide association scans. Bioinformatics 26, 2190–2191 (2010).
62. Pruim, R.J. et al. LocusZoom: regional visualization of genome-wide association scan results. Bioinformatics 26, 2336–2337 (2010).
63. Delaneau, O., Zagury, J.F. & Marchini, J. Improved whole-chromosome phasing for disease and population genetic studies. Nat. Methods 10, 5–6 (2013).
64. Delaneau, O., Marchini, J. & Zagury, J.F. A linear complexity phasing method for thousands of genomes. Nat. Methods 9, 179–181 (2012).
65. Howie, B.N., Donnelly, P. & Marchini, J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS Genet. 5, e1000529 (2009).
66. Howie, B., Marchini, J. & Stephens, M. Genotype imputation with thousands of genomes. G3 (Bethesda) 1, 457–470 (2011).
67. Marchini, J. & Howie, B. Genotype imputation for genome-wide association studies. Nat. Rev. Genet. 11, 499–511 (2010).
Corrigendum: Genome-wide association study identifies multiple susceptibility loci for pulmonary fibrosis

Tasha E Fingerlin, Elissa Murphy, Weiming Zhang, Anna L Peljto, Kevin K Brown, Mark P Steele, James E Loyd, Gregory P Cosgrove, David Lynch, Steve Groshong, Harold R Collard, Paul J Wolters, Williamson Z Bradford, Karl Kossen, Scott D Seiwert, Roland M du Bois, Christine Kim Garcia, Megan S Devine, Gunnar Gudmundsson, Helgi J Isaksson, Naftali Kaminski, Yingze Zhang, Kevin F Gibson, Lisa H Lancaster, Joy D Cogan, Wendi R Mason, Toby M Maher, Philip L Molyneaux, Athol U Wells, Miriam F Moffatt, Moises Selman, Annie Pardo, Dong Soon Kim, James D Crapo, Barry J Make, Elizabeth A Regan, Dinesha S Walek, Jerry J Daniel, Yoichiro Kamatani, Diana Zelenika, Keith Smith, David McKean, Brent S Pedersen, Janet Talbert, Raven N Kidd, Cheryl R Markin, Kenneth B Beckman, Mark Lathrop, Marvin I Schwarz & David A Schwartz

Nat. Genet. 45, 613–620 (2013); published online 14 April 2013; corrected after print 1 October 2013

In the version of this article initially published, the minor alleles for two SNPs (rs2076295 and rs7934606) were listed incorrectly in Table 1. The correct minor allele for rs2076295 is G, and the correct minor allele for rs7934606 is T. As a result, there were two incorrect statements in the text regarding the relationship between the disease risk allele of rs2076295 and DSP expression. The text should have stated that the allele of rs2076295 associated with increased risk of pulmonary fibrosis is associated with decreased DSP expression rather than increased DSP expression. These errors have been corrected in the HTML and PDF versions of the article.