RHEUMATOLOGY

Original article

Associations of the *MUC5B* promoter variant with timing of interstitial lung disease and rheumatoid arthritis onset

Gregory McDermott [1,2], Ritu Gill[2,3], Staci Gagne[2,4], Suzanne Byrne[2,4],
Weixing Huang[1], Jing Cui [1,2], Lauren Prisco[1], Alessandra Zaccardelli[1],
Lily Martin[1], Vanessa L. Kronzer [5], Matthew Moll[2,6,7], Michael H. Cho[2,6,7],
Nancy Shadick[1,2], Paul F. Dellaripa[1,2], Tracy Doyle[2,6] and Jeffrey A. Sparks [1,2]

Abstract

Objectives. To investigate the associations of the common *MUC5B* promoter variant with timing of RA-associated interstitial lung disease (RA-ILD) and RA onset.

Methods. We identified patients with RA meeting 2010 ACR/EULAR criteria and available genotype information in the Mass General Brigham Biobank, a multihospital biospecimen and clinical data collection research study. We determined RA-ILD presence by reviewing all RA patients who had CT imaging, lung biopsy or autopsy results. We determined the dates of RA and RA-ILD diagnoses by manual records review. We examined the associations of the *MUC5B* promoter variant (G>T at rs35705950) with RA-ILD, RA-ILD occurring before or within 2 years of RA diagnosis and RA diagnosis at age >55 years. We used multivariable logistic regression to estimate odds ratios (ORs) for each outcome by *MUC5B* promoter variant status, adjusting for potential confounders including genetic ancestry and smoking.

Results. We identified 1005 RA patients with available genotype data for rs35705950 (mean age 45 years, 79% female, 81% European ancestry). The *MUC5B* promoter variant was present in 155 (15.4%) and was associated with RA-ILD [multivariable OR 3.34 (95% CI 1.97, 5.60)], RA-ILD before or within 2 years of RA diagnosis [OR 4.01 (95% CI 1.78, 8.80)] and RA onset after age 55 years [OR 1.52 (95% CI 1.08, 2.12)].

Conclusions. The common *MUC5B* promoter variant was associated with RA-ILD onset earlier in the RA disease course and older age of RA onset. These findings suggest that the *MUC5B* promoter variant may impact RA-ILD risk early in the RA disease course, particularly in patients with older-onset RA.

Key words: RA, genetics, respiratory, epidemiology, CT scanning

BASIC SCIENCE

Rheumatology key messages

- The *MUC5B* promoter variant (rs35705950) may increase RA-ILD risk in the early RA period.
- The *MUC5B* rs35705950 genotype was also associated with older age of RA onset.
- The lungs and aging may play a crucial role in the pathogenesis of both RA and RA-ILD.

---

[1]Division of Rheumatology, Inflammation, and Immunity, Brigham and Women's Hospital, [2]Department of Medicine, Harvard Medical School, [3]Department of Radiology, Beth Israel Deaconess Medical Center, [4]Department of Radiology, Brigham and Women's Hospital, Boston, MA, [5]Division of Rheumatology, Mayo Clinic, Rochester, MN, [6]Division of Pulmonary and Critical Care Medicine and [7]Channing Division of Network Medicine, Brigham and Women's Hospital, Boston, MA, USA

Correspondence to: Jeffrey A. Sparks, Division of Rheumatology, Inflammation, and Immunity, Brigham and Women's Hospital, 60 Fenwood Road, 6016U, Boston, MA 02115, USA.
E-mail: jsparks@bwh.harvard.edu

Introduction

RA-associated interstitial lung disease (RA-ILD) is a serious extra-articular disease manifestation of RA that results in increased mortality [1–4]. The common *MUC5B* promoter variant (G>T, rs35705950) is associated with increased mucin 5B production in distal airway secretory cells and increased risk of RA-ILD and idiopathic pulmonary fibrosis (IPF) [5, 6]. While this genetic marker is an established risk factor for RA-ILD, its impact on the timing of RA-ILD relative to RA diagnosis

and articular RA disease course has not been previously investigated.

Prior investigations have established the importance of the *MUC5B* promoter variant in both the general population and patients with RA. After initial genetic studies noted associations between the *MUC5B* rs35705950 genotype and IPF, subsequent investigations demonstrated enrichment of the same variant in RA-ILD patients, particularly those with the usual interstitial pneumonia (UIP) phenotype [5, 6]. More recent population studies noted the combination of RA and the *MUC5B* promoter variant confers up to 10 times increased lifetime risk for ILD compared with the general population, suggesting an interplay between RA disease, ageing and genetic factors [7]. Although the *MUC5B* promoter variant is an established risk factor for RA-ILD, its impact on the timing of RA-ILD and RA disease features remains unknown.

Our study of RA patients used a large institutional biobank to examine the association of the *MUC5B* promoter variant with timing of ILD and RA disease features. We hypothesized that the *MUC5B* promoter variant would contribute to RA-ILD onset earlier in the RA disease course.

Methods

Study population

We performed a retrospective cohort study using linked genomic and clinical information from the Mass General Brigham (MGB) Biobank, a large, multihospital biospecimen collection program that recruits patients from inpatient and outpatient sites throughout the MGB healthcare system in the greater Boston area [8]. A total of 123 018 patients enrolled between 1 January 2010 and February 2021 and consented to provide blood samples and linked electronic health records (EHRs). Genomic data were available for 32 516 patients.

Identification of patients with RA

We identified patients with RA in the MGB Biobank using a previously validated algorithm that incorporates diagnosis codes, laboratory results and natural language processing of EHR terms with 95% positive predictive value at 97% specificity for RA as defined by the 2010 ACR/EULAR classification criteria [9, 10]. For the purposes of this study investigating the *MUC5B* promoter variant and RA features, we limited our study to RA patients with available genotyping of the *MUC5B* promoter variant ($n = 1005$). Fig. 1 details the analysed study sample. We extracted EHR data including imaging reports, clinical notes, RF and anti-CCP status, pathology reports and autopsy results. We performed a medical records review to confirm that all patients met the 2010 ACR/EULAR RA classification criteria [11]. Since laboratory testing was obtained for clinical purposes, some patients lacked RF ($n = 56$) or anti-CCP ($n = 148$) results. The study was approved by the MGB

Institutional Review Board (protocol 2019P000264) and adheres to the Declaration of Helsinki. All patients provided written informed consent at MGB Biobank enrolment.

Exposure: *MUC5B* promoter variant status

The exposure of interest was the common *MUC5B* promoter variant (G>T, rs35705950). We defined each patient's genotype as either absent (GG), one copy (GT) or two copies (TT). The MGB Biobank performed genotyping using the Illumina Multi-Ethnic Genotyping Array (MEGA), Expanded Multi-Ethnic Genotyping Array (MEGA^{EX}) and Multi-Ethnic Global (MEG) chips (Illumina, San Diego, CA, USA) and curated data using PLINK [12]. Imputation methods used to obtain relevant single-nucleotide polymorphisms (SNPs), including *MUC5B* rs35705950 status and RA genetic risk score, have been previously described [13]. Briefly, imputation was performed on the Michigan Imputation Server with the Minimac3 imputation algorithm using the 1000 Genomes Project phase 3 version 5 as the reference. Quality control required maximum per SNP missing of <0.02, maximum per person missing of <0.02, a Hardy–Weinberg disequilibrium *P*-value of 0.00001 and imputation quality ≥0.8. There were 55 patients who had rs35705950 directly genotyped using a TaqMan genotyping assay (Applied Biosystems, Waltham, MA, USA). There was 100% concordance between imputed and direct genotyping for these patients.

Outcomes: RA features

RA-ILD

We reviewed all patients with RA and available CT chest imaging ($n = 388$ with *MUC5B* genotype data) for clinically apparent RA-ILD. Based on previously published criteria, we considered patients to have clinically apparent RA-ILD when they had CT chest findings consistent with ILD as well as a physician's clinical diagnosis of RA-ILD [4]. We considered all other patients, including those without CT chest imaging ($n = 617$), those without evidence of ILD ($n = 308$) and those with CT chest findings that were indeterminate for RA-ILD or those with CT chest findings but lacking a clinical RA-ILD diagnosis ($n = 28$) to not have ILD in the main analyses. We determined the radiologic ILD subtype based on the radiologist's impression recorded in the clinical CT imaging reports.

RA-ILD before or within 2 years of RA diagnosis

To study the temporal relationship between RA and RA-ILD, we identified the dates of each diagnosis by manual medical records review. The ILD diagnosis date was the date of an abnormal imaging finding or the date of clinical suspicion for RA-ILD prompting further testing with imaging or pulmonary function testing, whichever was earlier. We determined the first documented date of RA diagnosis meeting 2010 ACR/EULAR classification criteria by medical records review. We identified patients diagnosed with RA-ILD before or within 2 years of RA

Fɪɢ. 1 Study sample
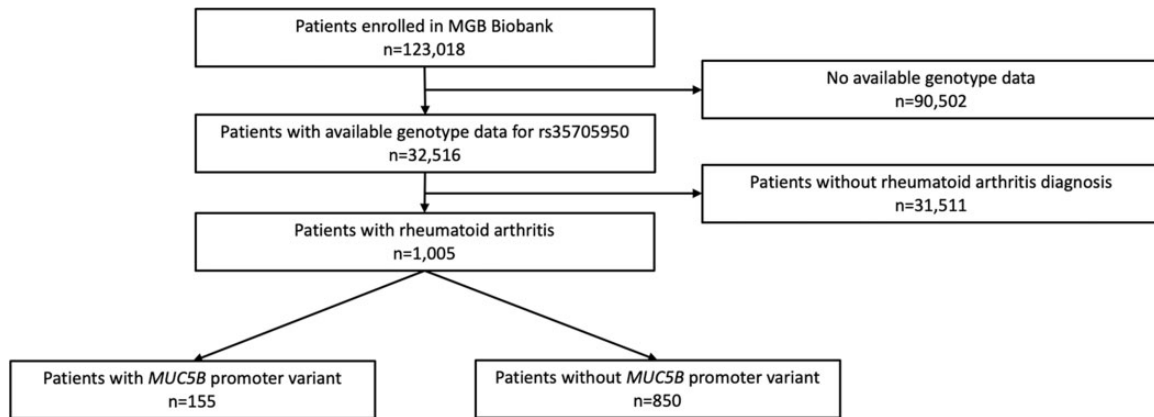
diagnosis, a timeframe previously used in the study of the early RA period [14].

Age at RA diagnosis

To examine the association of the *MUC5B* promoter variant with the timing of RA diagnosis, we investigated the age of each patient at RA diagnosis. In our main analysis, we dichotomized the RA diagnosis before and after age 55 years. This age coincides with the median age of RA diagnosis and is consistent with prior studies investigating the age of RA onset [15–18]. We also performed a sensitivity analysis using an age cut-off of 60 years.

Covariates

We assessed several heritable covariates that may influence RA and lung disease due to population stratification. To minimize the risk of confounding by ancestry, we determined the principal components of ancestry and defined European Ancestry using the first three principal components for the Utah Residents with Northern and Western European ancestry (CEU) population, using the 1000 Genomes Project as a reference.

We also calculated a genetic risk score for RA (RA-GRS) weighted by the natural log of effect size from previous large transethnic studies including 101 non-*HLA* SNPs and amino acid haplotypes of *HLA* [19]. Amino acid haplotypes of *HLA* were determined using the SNP2HLA software package [20].

We included smoking status, a risk factor for ILD and RA, as well as sex and age at RA diagnosis as covariates [21–23]. Smoking status and pack-years of cigarette exposure were obtained from health questionnaires completed at MGB Biobank enrolment and supplemented by medical records review for patients who did not complete the questionnaire. There were no missing data for exposures or covariates in the analyses.

Statistical analysis

We assessed characteristics at the time of enrolment in the MGB Biobank for RA patients with and without the *MUC5B* promoter variant. We used Fisher's exact test to compare proportions and Wilcoxon rank sum test with medians and interquartile ranges (IQRs) for continuous variables.

To examine the association between *MUC5B* promoter variant genotype and RA-ILD, timing of RA-ILD and age of RA diagnosis, we performed logistic regression to calculate odds ratios (ORs) and 95% CIs per copy of rs35705950 G>T. The dependent variables were the three binary outcomes: ever RA-ILD, RA-ILD before or within 2 years of RA diagnosis and dichotomized age at RA diagnosis in separate analyses. Base models were unadjusted. The multivariable model adjusted for potential confounders including age at RA diagnosis, sex, RA-GRS and genetic ancestry as well as smoking status (never/past/current) and pack-years of smoking (as a continuous variable) to account for the possible confounding/mediating effect that smoking may play in *MUC5B* status and RA and RA-ILD risk [24].

We performed several sensitivity analyses to assess the robustness of our findings. These included analyses removing patients indeterminate for RA-ILD ($n = 28$) and restricting the study population to patients of European ancestry to further guard against population stratification. We also performed analyses limited to patients with CT chest imaging and in the subgroups of patients with and without RA-ILD. We also repeated our main analysis treating the *MUC5B* genotype as a categorical variable (GG, GT or TT).

Given the importance of smoking as a respiratory risk factor as well as observed differences in RA and RA-ILD incidence by sex, we performed analyses stratified by sex and smoking status. We included an interaction term between sex and *MUC5B* promoter variant status for each RA feature outcome and reported *P*-values obtained using the Wald test. We performed a similar analysis for the interaction between smoking status and

MUC5B promoter variant status. Analyses used R version 3.6.1 (R Foundation for Statistical Computing, Vienna, Austria) or SAS version 9.4 (SAS Institute, Cary, NC, USA). A two-sided *P*-value <0.05 was considered statistically significant, without adjustment for multiple comparisons.

Results

Patient characteristics

Of 32 516 patients with available genotype data for rs35705950, we identified 1005 patients with RA. The baseline characteristics of the RA patients are listed in Table 1. A total of 10 patients had two copies of the *MUC5B* promoter variant (TT genotype) and 145 had one copy (GT genotype). Overall, 15.4% of RA patients had at least one copy of the *MUC5B* promoter variant, compared with 15.8% of the entire MGB Biobank. There was no association between the rs35705950 genotype and the presence of RA in the MGB Biobank (*P* = 0.68). The prevalence of clinically apparent RA-ILD was 5.2% (52/1005) when using stringent RA-ILD criteria and 8.0% (80/1005) when

indeterminate patients were included, consistent with the prevalence of clinically apparent RA-ILD of 4.6–7.8% reported in other large US cohorts [2, 4, 25].

A greater proportion of RA patients with the *MUC5B* promoter variant had European ancestry (90% TT genotype, 93.1% GT genotype, 78.4% GG genotype; *P* < 0.0001), RA diagnosis at age >55 years (60% TT genotype, 37.9% GT genotype, 28.5% GG genotype; *P* = 0.008) and RA-ILD (30% TT genotype, 11.7% GT genotype, 3.8% GG genotype; *P* < 0.0001) compared with RA patients without the variant. Those with the *MUC5B* promoter variant were older at RA diagnosis (median 59.5 years TT genotype, 48.0 years GT genotype, 45.7 years GG genotype; *P* = 0.03). There were no significant differences between RA-GRS (11.7 TT genotype, 12.1 GT genotype, 12.1 GG genotype; *P* = 0.43) or seropositivity (50% TT genotype, 55.2% GT genotype, 61.2% GG genotype; *P* = 0.40) by the *MUC5B* genotype. Among patients with available autoantibody testing, 605/962 (62.8%) were seropositive, similar to other RA cohorts [15, 26–28]. The breakdown of serostatus among patients with available autoantibody information is provided in Supplementary Table S1, available at *Rheumatology* online. The timing and age of RA and RA-ILD in the subgroup

TABLE 1 Characteristics of patients with RA enrolled in the MGB Biobank by *MUC5B* promoter variant genotype (*n* = 1005)

Characteristics	*MUC5B* TT genotype (*n* = 10)	*MUC5B* GT genotype (*n* = 145)	*MUC5B* GG genotype (*n* = 850)	*P*-value
Demographics				
Sex, female, *n* (%)	8 (80)	111 (76.6)	676 (79.5)	0.71
European ancestry, *n* (%)	9 (90)	135 (93.1)	666 (78.4)	**<0.0001**
Smoking status, n (%)				0.21
Never	5 (50)	62 (42.8)	445 (52.4)	
Past	4 (40)	68 (46.9)	342 (40.2)	
Current	1 (10)	15 (10.3)	63 (7.4)	
Smoking pack-years, median (IQR)	0.13 (0–15.8)	2.5 (0–17.5)	0 (0–10)	0.09
RA clinical features				
Age at RA diagnosis, years, median (IQR)	59.5 (53.8–66.2)	48.0 (36.1–58.8)	45.7 (32.9–56.6)	**0.03**
RA diagnosis at age >55 years, *n* (%)	6 (60)	55 (37.9)	242 (28.5)	**0.008**
Seropositive RA, *n* (%)	5 (50)	80 (55.2)	520 (61.2)	0.40
RF status, *n* (%)				0.13
Positive	5 (50)	64 (44.1)	463 (54.5)	
Negative	4 (40)	73 (50.3)	340 (40.0)	
Missing	1 (10)	8 (5.5)	47 (5.5)	
Anti-CCP status, *n* (%)				0.76
Positive	4 (40)	48 (40)	376 (44.2)	
Negative	4 (40)	67 (46.2)	348 (40.9)	
Missing	2 (20)	20 (13.8)	126 (14.8)	
Shared epitope present, *n* (%)	4 (40)	71 (49.0)	375 (44.1)	0.34
RA-GRS, median (IQR)	11.7 (10.7–12.1)	12.1 (10.5–13.6)	12.1 (10.7–13.4)	0.43
RA-ILD, *n* (%)	3 (30)	17 (11.7)	32 (3.8)	**<0.0001**
UIP pattern	2 (20.0)	6 (4.1)	9 (1.1)	**0.0006**
Fibrotic NSIP pattern	0 (0)	5 (3.5)	6 (0.7)	**0.02**
Other ILD	0 (0)	2 (1.4)	9 (1.1)	0.70
Unknown/unclassified ILD	1 (10)	4 (2.8)	8 (0.9)	**0.02**

NSIP: non-specific interstitial pneumonia; UIP: usual interstitial pneumonia; GRS: genetic risk score. Bold denotes results that had *P*-value < 0.05.

TABLE 2 Associations of the *MUC5B* promoter variant with RA and RA-ILD characteristics

Outcomes	*MUC5B* promoter variant (GT or TT genotype) (*n* = 155)	No *MUC5B* promoter variant (*n* = 850)
RA-ILD, *n* (%)	20 (12.9)	32 (3.8)
Unadjusted OR[a] (95% CI)	**3.36 (2.02, 5.47)**	Ref
Adjusted OR[a,b] (95% CI)	**3.34 (1.97, 5.60)**	Ref
RA-ILD before or within 2 years of RA diagnosis, *n* (%)	9 (5.8)	7 (0.8)
Unadjusted OR[a] (95% CI)	**4.91 (2.23, 10.40)**	Ref
Adjusted OR[a,b] (95% CI)	**4.01 (1.78, 8.80)**	Ref
RA diagnosis at age >55 years, *n* (%)	61 (39.4)	242 (28.5)
Unadjusted OR[a] (95% CI)	**1.62 (1.18, 2.24)**	Ref
Adjusted OR[a,c] (95% CI)	**1.52 (1.08, 2.12)**	Ref

[a]OR per copy of rs35705950 G>T [GG genotype has zero copies (reference group), GT genotype has one copy, TT genotype has two copies]. [b]Adjusted for age at RA diagnosis, sex, genetic ancestry, RA-GRS, smoking and continuous smoking pack-years. [c]Adjusted for sex, genetic ancestry, RA-GRS, smoking and continuous smoking pack-years. Bold results have statistically significant odds ratios that do not include 1.0.

of RA-ILD patients is listed in Supplementary Table S2, available at *Rheumatology* online.

Associations of the *MUC5B* promoter variant with RA and RA-ILD characteristics

RA-ILD

RA patients with the *MUC5B* promoter variant had increased odds of RA-ILD [unadjusted OR 3.36 (95% CI 2.02, 5.47)] per copy of rs35705950 G>T (Table 2). These findings persisted in the multivariable model adjusting for age at RA diagnosis, sex, genetic ancestry, RA-GRS, smoking status and pack-years, with an OR for RA-ILD of 3.34 (95% CI 1.97, 5.60) per copy of the *MUC5B* promoter variant.

RA-ILD before or within 2 years of RA diagnosis

Patients with the *MUC5B* promoter variant had higher odds of RA-ILD before or within 2 years of RA diagnosis [OR 4.91 (95% CI 2.23, 10.40) per copy of rs35705950 G>T]. After multivariable adjustment, the OR was 4.01 (95% CI 1.78, 8.80). For patients with the *MUC5B* promoter variant, the odds of RA-ILD before or within 2 years were higher than the odds of RA-ILD later in the disease course [OR 2.62 (95% CI 1.33, 4.87) for RA-ILD after 2 years per copy of rs35705950 G>T. Furthermore, when we limited to patients with RA-ILD, 9/20 (45%) with the *MUC5B* promoter variant (GT or TT genotype) had RA-ILD within 2 years of RA diagnosis compared with 7/32 (22%) patients without the promoter variant (*P* = 0.078).

RA diagnosis after age 55 years

The *MUC5B* promoter variant was associated with increased odds of RA diagnosis after age 55 years [OR 1.63 (95% CI 1.14, 2.32) per copy of rs35705950 G>T]. The unadjusted OR for RA diagnosis after age 55 years was 1.80 (95% CI 1.24, 2.60) per copy of the *MUC5B* promoter variant. The OR was 1.67 (95% CI 1.13, 2.46) in the multivariable model. When we removed 52 patients with RA-ILD and 28 patients indeterminate for

RA-ILD from our analysis, we observed a similar trend towards an older age of RA diagnosis that did not meet statistical significance [multivariable OR 1.44 (95% CI 0.998, 2.09) for RA diagnosis after age 55 years per copy of rs35705950 G>T] (Supplementary Tables S3 and S4, available at *Rheumatology* online).

Analyses stratified by sex and smoking status

We did not observe a statistically significant interaction between *MUC5B* promoter variant status and sex in stratified analyses (Table 3). Among females, there was a statistical association of the *MUC5B* promoter variant with older-onset RA [multivariable OR 1.67 for RA at >55 years (95% CI 1.13, 2.46) per copy of rs35705950 G>T]. This contrasted with no association of the *MUC5B* promoter variant among men for older-onset RA [multivariable OR 1.11 (95% CI 0.55, 2.20)]. However, the *P*-value for interaction by sex remained above the threshold for statistical significance (*P* for interaction = 0.35).

Similarly, we found no evidence for interaction of *MUC5B* and smoking status for any of the outcomes examined (Table 4).

Sensitivity analyses

We performed several sensitivity analyses. We excluded 28 patients indeterminate for clinically apparent RA-ILD due to non-specific CT chest abnormalities or lacking clinical evidence of RA-ILD and repeated our main analysis (Supplementary Table S5, available at *Rheumatology* online). We also performed an analysis excluding 195 patients with non-European ancestry (Supplementary Table S6, available at *Rheumatology* online). As CT chest imaging was obtained for clinical indications rather than universally on all patients, we performed an analysis restricted to patients who had available CT imaging (*n* = 388) (Supplementary Tables S7 and S8, available at *Rheumatology* online). We investigated the association of the *MUC5B* promoter variant

TABLE 3 Associations of the *MUC5B* promoter variant with RA and RA-ILD characteristics stratified by sex

Outcomes	Female (*n* = 795)		Male (*n* = 210)		*P* for interaction
	MUC5B promoter variant GT or TT genotype (*n* = 119)	No *MUC5B* promoter variant (*n* = 676)	*MUC5B* promoter variant GT or TT genotype (*n* = 36)	No *MUC5B* promoter variant (*n* = 174)	
RA-ILD, *n* (%)	13 (10.9)	21 (3.1)	7 (19.4)	11 (6.3)	
Unadjusted OR[a] (95% CI)	**3.34 (1.79, 5.99)**	1.00 (ref)	**3.39 (1.33, 8.37)**	1.00 (ref)	0.98
Adjusted OR[a,b] (95% CI)	**3.31 (1.72, 6.17)**	1.00 (ref)	**4.13 (1.51, 11.34)**	1.00 (ref)	0.82
RA-ILD before or within 2 years of RA diagnosis, *n* (%)	5 (4.2)	4 (0.6)	4 (11.1)	3 (1.7)	
Unadjusted OR[a] (95% CI)	**4.13 (1.36, 10.94)**	1.00 (ref)	**6.57 (1.84, 25.35)**	1.00 (ref)	0.57
Adjusted OR[a,b] (95% CI)	**3.52 (1.13, 9.93)**	1.00 (ref)	**9.20 (2.19, 46.68)**	1.00 (ref)	0.40
RA diagnosis at age >55 years, *n* (%)	45 (37.8)	165 (24.4)	16 (44.4)	77 (44.3)	
Unadjusted OR[a] (95% CI)	**1.80 (1.24, 2.60)**	1.00 (ref)	1.14 (0.58, 2.23)	1.00 (ref)	0.24
Adjusted OR[a,c] (95% CI)	**1.67 (1.13, 2.46)**	1.00 (ref)	1.11 (0.55, 2.20)	1.00 (ref)	0.35

[a]OR per copy of rs35705950 G>T [GG genotype has zero copies (reference group), GT genotype has one copy, TT genotype has two copies]. [b]Adjusted for age at RA diagnosis, sex, genetic ancestry, RA-GRS, smoking and continuous smoking pack-years. [c]Adjusted for sex, genetic ancestry, RA-GRS, smoking and continuous smoking pack-years. Bold results have statistically significant odds ratios that do not include 1.0.

TABLE 4 Associations of the *MUC5B* promoter variant with RA and RA-ILD characteristics stratified by smoking status

Outcomes	Ever smokers (*n* = 493)		Never smokers (*n* = 512)		*P* for interaction
	MUC5B promoter variant GT or TT genotype (*n* = 88)	No *MUC5B* promoter variant (*n* = 405)	*MUC5B* promoter variant GT or TT genotype (*n* = 67)	No *MUC5B* promoter variant (*n* = 445)	
RA-ILD, *n* (%)	14 (15.9)	22 (5.4)	6 (9.0)	10 (2.2)	
Unadjusted OR[a] (95% CI)	**3.11 (1.64, 5.74)**	1.00 (ref)	**3.55 (1.43, 8.02)**	1.00 (ref)	0.80
Adjusted OR[a,b] (95% CI)	**3.35 (1.73, 6.40)**	1.00 (ref)	**3.51 (1.39, 8.17)**	1.00 (ref)	0.92
RA-ILD before or within 2 years of RA diagnosis, *n* (%)	5 (5.7)	5 (1.2)	4 (6.0)	2 (0.4)	
Unadjusted OR[a] (95% CI)	**4.29 (1.48, 11.55)**	1.00 (ref)	**5.73 (1.60, 18.54)**	1.00 (ref)	0.71
Adjusted OR[a,b] (95% CI)	**4.03 (1.34, 11.83)**	1.00 (ref)	**4.90 (1.31, 16.41)**	1.00 (ref)	0.84
RA diagnosis at age >55 years, *n* (%)	40 (45.5)	139 (34.3)	21 (31.3)	103 (23.1)	
Unadjusted OR[a] (95% CI)	**1.57 (1.02, 2.42)**	1.00 (ref)	1.57 (0.94, 2.56)	1.00 (ref)	0.99
Adjusted OR[a,c] (95% CI)	1.54 (0.99, 2.39)	1.00 (ref)	1.42 (0.84, 2.36)	1.00 (ref)	0.82

[a]OR per copy of rs35705950 G>T [GG genotype has zero copies (reference group), GT genotype has one copy, TT genotype has two copies). [b]Adjusted for age at RA diagnosis, sex, genetic ancestry and RA-GRS. [c]Adjusted for sex, genetic ancestry and RA-GRS. Bold results have statistically significant odds ratios that do not include 1.0.

with RA diagnosis after age 60 years rather than age 55 years (Supplementary Table S9, available at *Rheumatology* online). We also repeated our main analysis considering genotype (TT, GT or GG) as a categorical variable (Supplementary Table S10, available at *Rheumatology* online). Finally, to account for the possibility of misclassification of smoking status due to changes that occurred between RA diagnosis and Biobank enrolment, we examined differences in ever *vs* never smokers (Supplementary Table S11, available at *Rheumatology* online). Our main findings remained unchanged in these analyses.

Discussion

In this study using a large institutional biobank, the common *MUC5B* promoter variant was associated with RA-ILD, older age of RA onset and earlier timing of RA-ILD. We discovered a strong association between the *MUC5B* rs35705950 genotype and RA-ILD onset before or within 2 years of RA diagnosis, even after adjusting for smoking status, serostatus, demographics and other genetic factors. Furthermore, the *MUC5B* promoter variant was associated with older age of RA diagnosis and a higher proportion of patients diagnosed with RA after

age 55 years. These findings expand the understanding of the importance of the *MUC5B* gene in RA and RA-ILD and suggest a possible pathogenic role for lung inflammation in the development of articular RA. Considering that the *MUC5B* promoter variant is relatively common, future studies should evaluate its utility for RA-ILD risk stratification in combination with other risk factors in specific populations, such as those with older age of RA onset or early RA.

This study adds to the growing literature on the interplay between the *MUC5B* promoter variant and RA-ILD [29–32]. Seibold *et al.* [6] first described the *MUC5B* promoter variant as an IPF risk factor in 2011. Subsequent studies have demonstrated enrichment of the same variant in RA-ILD patients, especially those with the UIP pattern, which shares many risk factors and histopathologic features with IPF [5]. The importance of the rs35705950 genotype in RA-ILD has been confirmed in recent investigations, including a Finnish study that found a 10-fold increased risk of ILD in patients with RA and the *MUC5B* promoter variant compared with the general population [7]. Despite other studies that variably describe seropositive RA as a RA-ILD risk factor, we did not observe significant differences in serostatus related to the *MUC5B* genotype [5, 25, 33]. This finding requires further investigation, as it may inform the understanding of mechanistic pathways linking RA, autoantibody production and lung inflammation. To our knowledge, this is the first study investigating the impact of the rs35705950 genotype on RA-ILD timing within the RA disease course.

Our findings also expand the understanding of the complex relationship between ageing, RA and ILD [34]. A recent study of lifetime RA-ILD risk noted a rapid increase in RA-ILD after age 65 years among those with the *MUC5B* promoter variant, which may relate to our finding of increased age of RA diagnosis among patients with the variant [7]. Similarly, a recent Swedish investigation of *MUC5B* and other ILD risk genes noted that older age of RA diagnosis was associated with RA-ILD, although the specific interaction of this finding with the *MUC5B* promoter variant was not reported [35]. Further evidence for an interaction between ageing, RA and ILD comes from the association between shorter telomeres and RA as well as IPF [36–39]. Interestingly, our study found a trend towards older age of RA onset with the *MUC5B* promoter variant, even among patients without RA-ILD. This finding will need further replication in studies with universal screening for RA-ILD. In our study, the association of the *MUC5B* promoter variant with older-onset RA was statistically significant among women but not men. However, the sample size may have been limited to find a significant interaction by sex. Future studies should investigate the relationship between *MUC5B* and RA onset after menopause, given previous associations of postmenopausal factors with RA risk and features [40–42].

The strengths of our study include the use of detailed clinical data to identify RA features and the incorporation of additional genetic data beyond the rs35705950 SNP, allowing for analyses adjusted for genetic ancestry, shared epitope status and RA-GRS. We confirmed RA and RA-ILD diagnoses using stringent criteria, verified the RA diagnosis through detailed medical records review and identified RA-ILD using clinical notes, imaging reports, autopsies and pathology findings. We included established RA-ILD risk factors including smoking status and pack-years [43]. These findings further characterize the impact of the most important known RA-ILD risk factor.

Our study has certain limitations. The cohort included mostly patients of European genetic ancestry, which may limit generalizability to other populations. However, we did not restrict the analysis to patients with European ancestry, unlike other studies [7, 35]. Furthermore, the association of the *MUC5B* promoter variant with RA-ILD has been established in a multiethnic cohort [5]. We relied on clinical data to identify RA-ILD, which raises the possibility of undiagnosed RA-ILD among patients who did not have imaging performed. However, we classified all patients without available CT chest imaging as 'no ILD', so misclassification of these patients would be expected to bias our RA-ILD findings towards the null and would not explain the associations that we reported [44]. In a sensitivity analysis, we removed patients who were indeterminate for RA-ILD from the analysis and found similar results. Furthermore, as the focus of our study was on clinically apparent RA-ILD, we were unable to assess whether our findings extend to patients with subclinical RA-ILD. Future studies to confirm our results in asymptomatic RA patients screened by CT scan are needed.

We examined many important covariates, including smoking, RA-related autoantibodies and RA serostatus, but residual confounding remains possible from factors such as inhalants, pollution, disease activity and RA treatments [26]. However, the RA-specific factors would not be relevant for analyses of incident RA and would mediate, not confound, any relationships between genetic factors and RA-ILD risk. We collected detailed information about smoking status from health questionnaires at Biobank enrolment. However, this time point may not coincide with RA or RA-ILD diagnosis and therefore misclassification of smoking status is possible. We did not adjust for multiple comparison testing, although our outcomes were pre-specified based on possible pathogenic mechanisms linking *MUC5B* to earlier-onset RA-ILD. Since all patients had RA, there is a possibility of index event bias, which has been observed in other studies of outcomes related to genetic factors, including prognosis of ILD patients with *MUC5B* [45]. However, our main inclusion criterion was RA diagnosis, which was not associated with the *MUC5B* promoter variant in our dataset and has not demonstrated a consistent and strong association with RA in RA genome-wide association studies, although was modestly associated with RA in a recent large population study [7, 19]. While RA-ILD subtypes were available on many patients, numbers were insufficient to restrict analyses to a particular subtype, such

as UIP, that may be more homogeneous. We focused our analysis on the *MUC5B* promoter variant, which is the strongest known genetic risk factor for RA-ILD. Future studies that incorporate additional ILD risk genes may further characterize the importance of genetic risk factors for RA-ILD. We primarily investigated factors present at diagnosis and did not have longitudinal data available on RA disease severity and activity, which have been linked to RA-ILD risk [26]. Additional studies are needed to characterize how these RA features interplay with *MUC5B* and RA-ILD subtypes. Finally, we did not have an available independent validation cohort, so our findings should be considered hypothesis-generating until replicated.

In conclusion, we found that the common *MUC5B* promoter variant was associated with RA-ILD, RA-ILD occurring before or within 2 years of RA diagnosis and RA onset after age 55 years. These findings confirm the importance of the *MUC5B* gene as an RA-ILD risk factor and suggest that it may impact RA-ILD risk early in the RA disease course, particularly in those with older-onset RA, regardless of serostatus. These findings may lead to the identification of RA patients amenable to screening or prevention strategies for RA-ILD.

Data availability statement

Data are available upon request and with appropriate institutional review board approval.

Supplementary data

Supplementary data are available at *Rheumatology* online.

References

1 Myasoedova E, Crowson CS, Turesson C, Gabriel SE, Matteson EL. Incidence of extraarticular rheumatoid arthritis in Olmsted County, Minnesota, in 1995–2007 versus 1985–1994: a population-based study. J Rheumatol 2011;38:983–9.

2 Sparks JA, Jin Y, Cho S-K *et al.* Prevalence, incidence and cause-specific mortality of rheumatoid arthritis-associated interstitial lung disease among older rheumatoid arthritis patients. Rheumatology (Oxford) 2021;60: 3689–98.

3 Hyldgaard C, Hilberg O, Pedersen AB *et al.* A population-based cohort study of rheumatoid arthritis-associated interstitial lung disease: comorbidity and mortality. Ann Rheum Dis 2017;76:1700–6.

4 Bongartz T, Nannini C, Medina-Velasquez YF *et al.* Incidence and mortality of interstitial lung disease in rheumatoid arthritis: a population-based study. Arthritis Rheum 2010;62:1583–91.

5 Juge P-A, Lee JS, Ebstein E *et al. MUC5B* promoter variant and rheumatoid arthritis with interstitial lung disease. N Engl J Med 2018;379:2209–19.

6 Seibold MA, Wise AL, Speer MC *et al.* A common *MUC5B* promoter polymorphism and pulmonary fibrosis. N Engl J Med 2011;364:1503–12.

7 Palomäki A, Palotie A, Koskela J *et al.* Lifetime risk of rheumatoid arthritis-associated interstitial lung disease in *MUC5B* mutation carriers. Ann Rheum Dis 2021;80: 1530–6.

8 Karlson EW, Boutin NT, Hoffnagle AG, Allen NL. Building the partners healthcare biobank at partners personalized medicine: informed consent, return of research results, recruitment lessons and operational considerations. J Pers Med 2016;6:2–11.

9 Liao KP, Cai T, Gainer V *et al.* Electronic medical records for discovery research in rheumatoid arthritis. Arthritis Care Res 2010;62:1120–7.

10 Huang S, Huang J, Cai T *et al.* Impact of ICD10 and secular changes on electronic medical record rheumatoid arthritis algorithms. Rheumatology (Oxford) 2020;59:3759–66.

11 Aletaha D, Neogi T, Silman AJ *et al.* 2010 rheumatoid arthritis classification criteria: an American College of Rheumatology/European League Against Rheumatism collaborative initiative. Arthritis Rheum 2010;62:2569–81.

12 Purcell S, Neale B, Todd-Brown K *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 2007;81:559–75.

13 Knevel R, Le Cessie S, Terao CC *et al.* Using genetics to prioritize diagnoses for rheumatology outpatients with inflammatory arthritis. Sci Transl Med 2020;12:eaay1548.

14 Raza K, Buckley CE, Salmon M, Buckley CD. Treating very early rheumatoid arthritis. Best Pract Res Clin Rheumatol 2006;20:849–63.

15 Sparks JA, O'Reilly ÉJ, Barbhaiya M *et al.* Association of fish intake and smoking with risk of rheumatoid arthritis and age of onset: a prospective cohort study. BMC Musculoskelet Disord 2019;20:2.

16 Barton JL, Schmajuk G, Trupin L *et al.* Poor knowledge of methotrexate associated with older age and limited English-language proficiency in a diverse rheumatoid arthritis cohort. Arthritis Res Ther 2013;15: R157.

17 Baghdadi LR, Woodman RJ, Shanahan EM, Mangoni AA. The impact of traditional cardiovascular risk factors on cardiovascular outcomes in patients with rheumatoid arthritis: a systematic review and meta-analysis. PLoS One 2015;10:e0117952.

18 Lu B, Hiraki LT, Sparks JA *et al.* Being overweight or obese and risk of developing rheumatoid arthritis among women: a prospective cohort study. Ann Rheum Dis 2014;73:1914–22.

19 Okada Y, Wu D, Trynka G *et al.* Genetics of rheumatoid arthritis contributes to biology and drug discovery. Nature 2014;506:376–81.

20 Jia X, Han B, Onengut-Gumuscu S *et al.* Imputing amino acid polymorphisms in human leukocyte antigens. PLoS One 2013;8:e64683.

21 Sparks JA, Chang SC, Deane KD *et al.* Associations of smoking and age with inflammatory joint signs among unaffected first-degree relatives of rheumatoid arthr patients: results from studies of the etiology of rheumatoid arthritis. Arthritis Rheumatol 2016;68: 1828–38.

22 Sugiyama D, Nishimura K, Tamaki K *et al.* Impact of smoking as a risk factor for developing rheumatoid arthritis: a meta-analysis of observational studies. Ann Rheum Dis 2010;69:70–81.

23 Baumgartner KB, Samet JM, Stidley CA, Colby TV, Waldron JA. Cigarette smoking: a risk factor for idiopathic pulmonary fibrosis. Am J Respir Crit Care Med 1997;155:242–8.

24 Baka Z, Buzás E, Nagy G. Rheumatoid arthritis and smoking: putting the pieces together. Arthritis Res Ther 2009;11:238.

25 Natalini JG, Baker JF, Singh N *et al.* Autoantibody seropositivity and risk for interstitial lung disease in a prospective male-predominant rheumatoid arthritis co-hort of U.S. veterans. Ann Am Thorac Soc 2021;18: 598–605.

26 Sparks JA, He X, Huang J *et al.* Rheumatoid arthritis disease activity predicting incident clinically apparent rheumatoid arthritis–associated interstitial lung disease: a prospective cohort study. Arthritis Rheumatol 2019;71: 1472–82.

27 Myasoedova E, Crowson CS, Kremers HM, Therneau TM, Gabriel SE. Is the incidence of rheumatoid arthritis rising? Results from Olmsted County, Minnesota, 1955–2007. Arthritis Rheum 2010;62:1576–82.

28 Solomon DH, Kremer J, Curtis JR *et al.* Explaining the cardiovascular risk associated with rheumatoid arthritis: traditional risk factors versus markers of rheumatoid arthritis severity. Ann Rheum Dis 2010;69:1920–5.

29 McDermott GC, Doyle TJ, Sparks JA. Interstitial lung disease throughout the rheumatoid arthritis disease course. Curr Opin Rheumatol 2021;33:284–91.

30 Sparks JA. Towards clinical significance of the *MUC5B* promoter variant and risk of rheumatoid arthritis-associated interstitial lung disease. Ann Rheum Dis 2021;60:3689–98.

31 Juge PA, Solomon JJ, van Moorsel CHM *et al. MUC5B* promoter variant rs35705950 and rheumatoid arthritis associated interstitial lung disease survival and progression. Semin Arthritis Rheum 2021;51:996–1004.

32 Wang N, Zhang Q, Jing X *et al.* The association between MUC5B mutations and clinical outcome in patients with rheumatoid arthritis-associated interstitial lung disease: a retrospective exploratory study in China. Med Sci Monit 2020;26:e920137.

33 Xie S, Li S, Chen B *et al.* Serum anti-citrullinated protein antibodies and rheumatoid factor increase the risk of rheumatoid arthritis-related interstitial lung disease: a meta-analysis. Clin Rheumatol 2021;40:4533–43.

34 Mohning MP, Amigues I, Demoruelle MK *et al.* Duration of rheumatoid arthritis and the risk of developing interstitial lung disease. ERJ Open Res 2021;7: 00633–2020.

35 Jönsson E, Ljung L, Norrman E *et al.* Pulmonary fibrosis in relation to genetic loci in an inception cohort of patients with early rheumatoid arthritis from northern Sweden. Rheumatology (Oxford) 2021;61:943–52.

36 Zeng Z, Zhang W, Qian Y *et al.* Association of telomere length with risk of rheumatoid arthritis: a meta-analysis and Mendelian randomization. Rheumatology (Oxford) 2020;59:940–7.

37 Stock CJW, Renzoni EA. Telomeres in interstitial lung disease. J Clin Med 2021;10:1384.

38 Armanios MY, Chen JJ-L, Cogan JD *et al.* Telomerase mutations in families with idiopathic pulmonary fibrosis. N Engl J Med 2007;356:1317–26.

39 Costenbader KH, Prescott J, Zee RY, De Vivo I. Immunosenescence and rheumatoid arthritis: does telomere shortening predict impending disease? Autoimmun Rev 2011;10:569–73.

40 Orellana C, Saevarsdottir S, Klareskog L *et al.* Postmenopausal hormone therapy and the risk of rheumatoid arthritis: results from the Swedish EIRA population-based case-control study. Eur J Epidemiol 2015;30:449–57.

41 Bengtsson C, Malspeis S, Orellana C *et al.* Association between menopausal factors and the risk of seronegative and seropositive rheumatoid arthritis: results from the Nurses' Health Studies. Arthritis Care Res 2017;69:1676–84.

42 Alpizar-Rodriguez D, Mueller RB, Möller B *et al.* Female hormonal factors and the development of anti-citrullinated protein antibodies in women at risk of rheumatoid arthritis. Rheumatology (Oxford) 2017;56:1579–85.

43 Kronzer VL, Huang W, Dellaripa PF *et al.* Lifestyle and clinical risk factors for incident rheumatoid arthritis-associated interstitial lung disease. J Rheumatol 2021; 48:656–63.

44 Rothman KJ, Greenland S, Lash TL. Modern epidemiology, 3rd ed. Philadelphia: Lippincott Williams & Wilkins, 2008.

45 Dudbridge F, Allen RJ, Sheehan NA *et al.* Adjustment for index event bias in genome-wide association studies of subsequent events. Nat Commun 2019;10:1561.