Online Data Supplement

Genome-Wide Association Study of Susceptibility to *Pseudomonas aeruginosa* Infection in Cystic Fibrosis

Authors

Boxi Lin, Jiafen Gong, Katherine Keenan, Fan Lin, Yu-Chung Lin, Julie Mésinèle, Claire Calmel, Badreddine Mohand Oumoussa, Pierre-Yves Boëlle, Loïc Guillot, Harriet Corvol, Valerie Waters, Lei Sun, and Lisa J Strug*

*Correspondence: lisa.strug@utoronto.ca

Supplementary Tables

Supplementary Table 1: Demographic and clinical characteristics of Canadian CF Gene Modifier Study (CGMS) GWAS sample and excluded samples.

Sample	**Chronic *Pa* age GWAS**[‡]	**Individuals without chronic *Pa* infection**[*]	**Individuals with chronic *Pa* but missing infection ages**[†]
N	1,037[‡]	546	834
Age[§]: Mean±SD (Year)	36.3±11.2	29.0±15.2	32.7±11.1
Age[§]: Range (Year)	(6.8, 71.1)	(3.54, 80.5)	(7.0, 78.9)
Sex: Male %	51.7%	59.3%	52.5%
Age at CF diagnosis: Mean ± SD (Year)	2.1±4.1	3.2±8.1	2.6±5.5
***CFTR* genotype distribution %,** **(**F508del/F508del, F508del/minimal function (Min), Min/Min**)**	(59.0%, 33.1%, 7.9%)	(57.7%, 33.2%, 9.2%)	(58.8%, 33.8%, 7.4%)
Sweat Chloride: Mean ± SD (ng/mL)	102.0±20.3	94.2±22.1	101.6±18.9
Newborn screening %	11.4%	14.0%	13.6%
First *Pa* age (Year)			
Mean ± SD	9.0±7.8	10.0±10.0	7.6±8.1
Range	(0.0, 52.1)	(0.1, 64.4)	(0.1, 60.1)
Range of diagnosis date	1966-11-06 ~ 2017-08-29	1968-09-10 ~ 2018-12-28	1970-06-25 ~ 2017-05-02
Missing rate %	7.9%	58.6%	42.6%
Chronic *Pa* age (Year)			
Mean ± SD	11.9±7.9	/	/
Range	(0.3, 52.1)	/	/
Range of diagnosis date	1966-11-06 ~ 2017-08-29	/	/

[‡] "Chronic *Pa* age GWAS" is the sample from CGMS for the GWAS of the chronic *Pa* age, where individuals have insufficient pancreatic function and non-missing chronic *Pa* age and are CF modulator-free.

[*] "Individuals without chronic *Pa* infection" individuals who were excluded from the GWAS as they were never chronically infected with *Pa* before the end of the follow-up.

[†] "Individuals without chronic *Pa* infection" individuals who were chronically infected with *Pa* before the end of the follow-up, but the date for defining chronic *Pa* was missing.

Supplementary Table 2: List of variants from the literature linked to CF lung phenotypes with corresponding p-values from the GWAS of chronic *Pa* age

Reference for studies	Variants	MAF in CGMS	Gene	Phenotype	*p*-values from GWAS for chronic *Pa* age
Emond, M.J., et al., *Exome sequencing of extreme phenotypes identifies DCTN4 as a modifier of chronic Pseudomonas aeruginosa infection in cystic fibrosis.*	rs11954652	0.06	*DCTN4*	Extreme Chronic *Pa* age, Exome sequencing, N<100	0.87
	rs8920	0.44	*CAV2*	Extreme Chronic *Pa* age, Exome sequencing, N=85+3,239	0.49
	rs34712518	0.06	*TMC6*	Extreme Chronic *Pa* age, Exome sequencing, N=85+3,239	0.07
Corvol, H., et al., *Genome-wide association meta-analysis identifies five modifier loci of lung disease severity in cystic fibrosis.*	rs3103933	0.30	*MUC4/MUC20*	SaKnorm, N=6,365	0.55
	rs57221529	0.22	*SLC9A3*	SaKnorm, N=6,365	0.19
	rs10742326	0.41	*EHF/APIP*	SaKnorm, N=6,365	0.99
	rs5952223	0.26	*AGTR2/SLC6A14*	SaKnorm, N=6,365	0.23

Supplementary Table 3. Results from first *Pa* age and chronic *Pa* age GWAS at the *AGTR2/SLC6A14* (chrXq22-q23).

SNP	Position	Minor allele	Major allele	Lung function Beta	Lung function p-value	First *Pa* Beta (Year)	First *Pa* p-value	Chronic *Pa* Beta (Year)	Chronic *Pa* p-value
rs7879546	115348275	C	T	0.073	1.11E-09	0.288	0.105	0.360	0.094
rs12559259	115348321	G	A	0.072	1.20E-09	0.222	0.213	0.442	0.041
rs12556757	115348290	A	G	0.072	1.43E-09	0.249	0.162	0.450	0.037
rs6520193	115365487	C	T	0.071	1.50E-09	0.272	0.125	0.329	0.126
rs12009976	115361495	A	G	0.071	1.65E-09	0.272	0.126	0.329	0.126
rs12559834	115348414	C	T	0.071	1.66E-09	0.285	0.108	0.352	0.101
rs4446858	115361808	C	T	0.071	1.69E-09	0.259	0.144	0.329	0.126
rs5952223	115386565	T	C	0.079	1.83E-09	0.446	0.028	0.294	0.232
rs7052638	115348603	C	A	0.072	1.84E-09	0.225	0.207	0.441	0.041
rs5905350	115356407	C	T	0.070	1.86E-09	0.301	0.091	0.362	0.092
rs5905340	115353303	C	T	0.070	2.02E-09	0.301	0.090	0.362	0.092
rs10854906	115359411	A	C	0.071	2.08E-09	0.196	0.273	0.419	0.052
rs4468049	115342921	G	C	0.070	2.16E-09	0.288	0.105	0.360	0.094
rs5905214	115359753	T	C	0.070	2.32E-09	0.196	0.273	0.419	0.052
rs5905282	115340298	T	C	0.070	2.33E-09	0.224	0.210	0.438	0.043

Comparison of effect sizes and p-values from GWASs of CF lung function [1], first *Pa* age (N=1,653) and chronic *Pa* age (N=1,037) at the CF lung function-associated loci *AGTR2/SLC6A14* (chrXq22-q23). 15 SNPs with smallest p-values reported from Corvol et al., 2015 are summarized and sorted here.

Supplementary Table 4. P-values in CF Lung Function GWAS (Corvol et al 2015) at variants identified from the first and chronic *Pa* age GWAS.

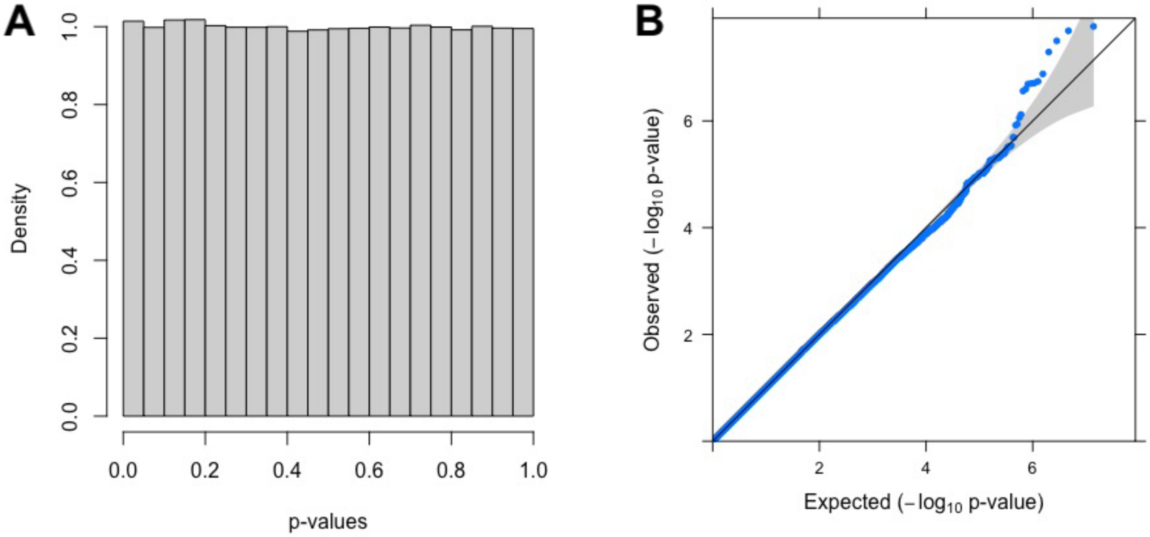
Variant Identified from *Pa* age GWAS	**Nearby gene**	**Chr**	**Position**	**MAF**	**P-value (EUR)**	**p-value (EUR+nonEUR)**	**P-value from Lung function GWAS**
First *Pa* age							
rs6847677	*EVC2*	4	5609053	0.04	$2.37\text{x}10^{-8}$	$4.73\text{x}10^{-8}$	0.79
rs60110289	*FAM47A*	X	34344790	0.04	$9.27\text{x}10^{-8}$	$5.41\text{x}10^{-7}$	0.05
Chronic *Pa* age							
rs62369766	*NNT / FGF10*	5	44229646	0.16	$7.13\text{x}10^{-8}$	$1.98\text{x}10^{-8}$	0.19
rs927553	*SPATA13*	13	24808505	0.41	$3.33\text{x}10^{-8}$	$1.91\text{x}10^{-8}$	0.43

Supplementary Table 5. Distributions of the genotypes for the two genome-wide significant SNPs across different sites

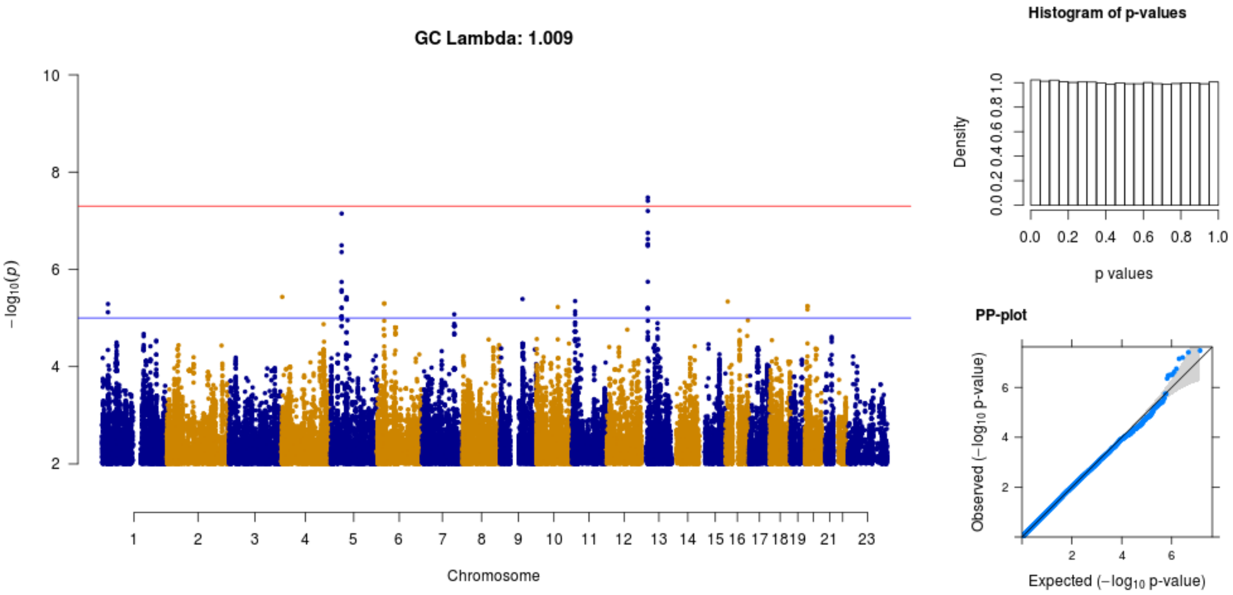
Center site	N	rs62369766 (*p*-value* = 0.080)						rs927553 (*p*-value* = 0.063)					
		G = 0	G = 1	G = 2	AF*	Beta (Year)	p	G = 0	G = 1	G = 2	AF*	Beta (Year)	p
OSK	155	120	34	1	0.116	-1.41	0.095	39	72	44	0.516	0.67	0.187
QDM	129	85	36	8	0.201	-0.27	0.774	40	69	20	0.420	1.61	0.060
OSM	94	68	24	2	0.149	-2.97	0.083	33	44	17	0.414	2.59	0.029
BSP	89	62	24	3	0.168	-4.5	3.76×10^{-4}	24	48	17	0.460	1.38	0.178
QSJ	62	43	17	2	0.169	-1.15	0.074	24	30	8	0.371	1.44	8.18×10^{-3}
AUA	58	42	13	3	0.163	-4.70	4.68×10^{-3}	26	24	8	0.344	-0.17	0.899
Other	450	299	143	8	0.176	-1.64	1.33×10^{-3}	138	232	80	0.436	1.35	4.23×10^{-4}
Aggregated	1,037	719	291	27	0.166	-2.00	1.98×10^{-8}	324	519	194	0.437	1.51	1.91×10^{-8}

Distribution of the two genome-wide significant SNPs associated with chronic *Pa* infection age by referring center, accompanied by site-specific effect size estimates and p-values. The GWAS cohort is derived from 35 Canadian sites. Referring sites with over 50 individuals are displayed in rows 1 to 6. Sites with fewer than 50 samples have been combined into the 'Other' category. Allele frequencies are based on the effect allele from the combined GWAS. P-values are calculated using Pearson’s Chi-square test, testing the null hypothesis that genotype distributions are consistent across the six major sites and the 'Other' category.

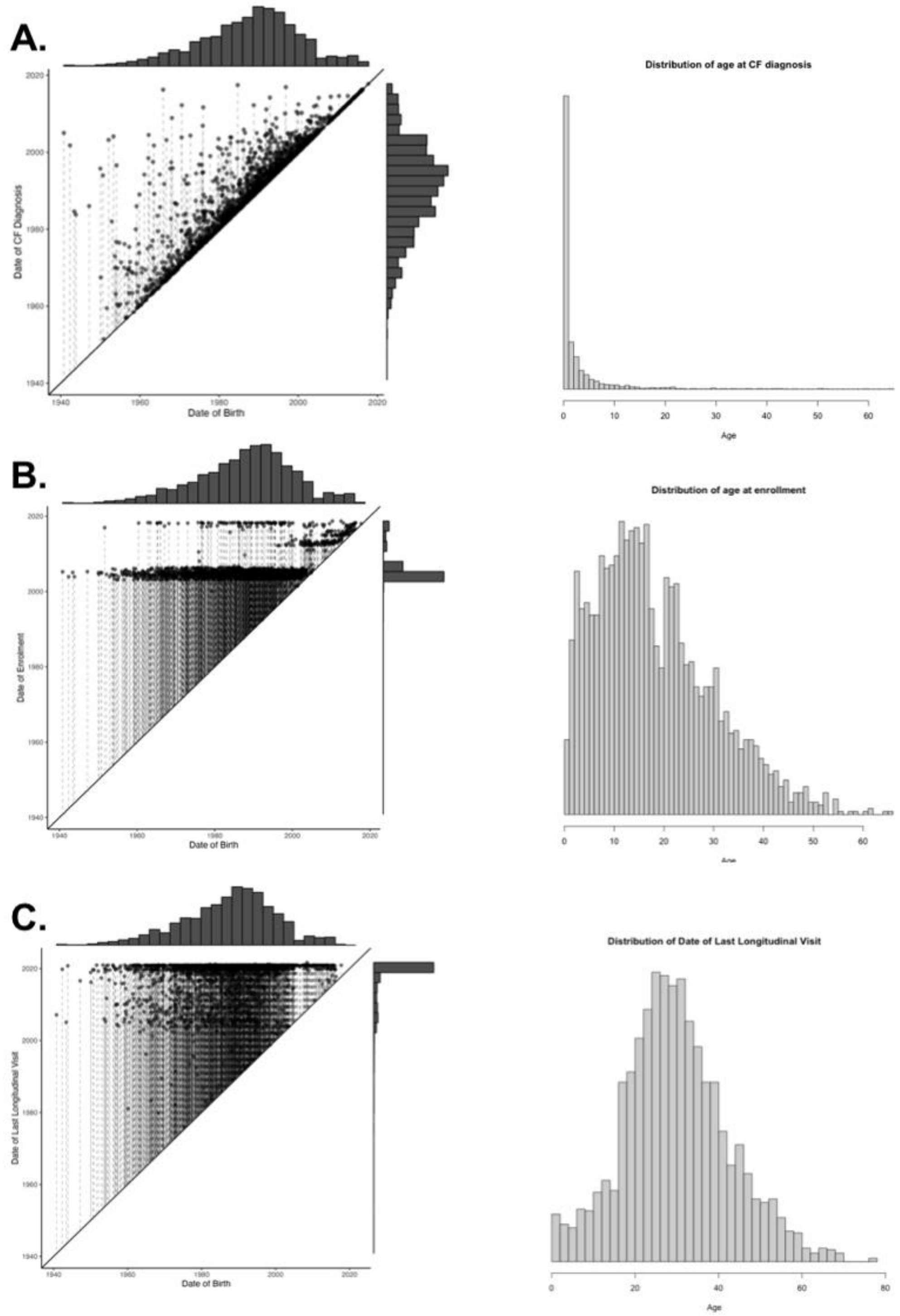
Supplementary Figures



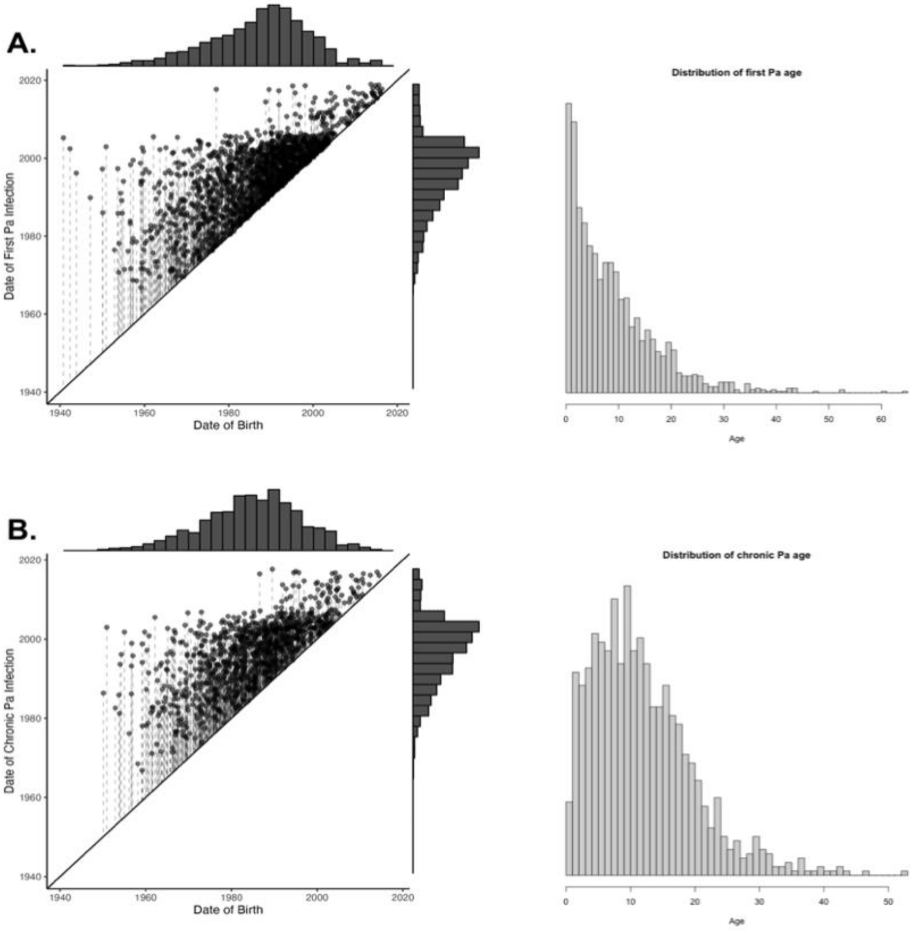
Supplementary Figure 1: The histogram (A) displays the distribution of p-values from GWAS of chronic *Pa* infection age in 1,037 individuals with CF, and the quantile-quantile plot (QQ-plot) (B) assesses the distribution of the p-values against the expected null distribution. In total, 6,994,213 SNPs were analyzed, all with MAF greater than 5%.



Supplementary Figure 2: The Manhattan, QQ-plot and histogram of the p-values for GWAS of chronic *Pa* age in N=991 European individuals. In total 6,994,213 SNPs were analyzed, all with MAF>5%. The red horizontal line corresponds to the genome-wide significance threshold of p-value=5×10^{-8}; the blue horizontal line corresponds to the suggestive level of p-value=1×10^{-5}.



Supplementary Figure 3: Characteristics of the Canadian CF Gene Modifier Study sample (N=2,728), including the distributions of their date of birth (x-axis of each plot on the left), diagnosis (y-axis in panel A, left), enrollment (y-axis in panel B, left), and last longitudinal visit (y-axis in panel C, left). The distance of each point to the diagonal line quantifies the age of that event for each individual, which is plotted as histograms on the right. Notice that the plots are based on individuals with non-missing data, so the samples used for plotting are slightly different across the three panels.



Supplementary Figure 4: Distribution of phenotypes for the Canadian CF Gene Modifier Study sample (N=2,728), including the distributions of their date of birth (x-axis of each plot on the left), first *Pa* infection (y-axis in panel A, left) and chronic *Pa* infection (y-axis in panel B, left). The distance of each point to the diagonal line quantifies the age at that event for an individual, which was plotted as histograms on the right. Notice that the plots are based on individuals with non-missing data, so the samples used for plotting are slightly different across the two panels.

Supplementary Appendix 1

Online Methods

Genotyping and SNP quality control (QC)

Genotyping, variant calling, and quality control (QC) were conducted in an earlier study of meconium ileus in CF [2]. To summarize, genotyping was performed on four Illumina platforms, the 610Quad, CNV370, 660W and Omni5. GenomeStudio V2011.1 was used for genotype calling, and SNP position and annotation information were based on Genome Reference Consortium 37 (GRCh37). We used PLINK 1.90b3x (Web Resources) to conduct QC. Major procedures included removing SNPs not annotated to chromosomes 1–22 or the X chromosome, those without an rs number, those with a call rate of less than 90%, duplicated variants, X chromosomal SNPs with a heterozygosity rate of more than 10% in males [3], SNPs that failed the Hardy-Weinberg Equilibrium test with a p-value of less than 5×10^{-30} and rare variants with a minor allele frequency of less than 1%.

Sample quality control

From the 2,741 genotyped samples in the Canadian Gene Modifier Study (CGMS), we: A) removed two individuals whose clinical records could not be verified, possibly due to a lack of longitudinal data or recent updates; B) excluded eight samples with mismatched sex, as determined by a discrepancy between the registry data and the genetically inferred sex through the X chromosome homozygosity rate (homozygosity estimate threshold was set at greater than 0.8 for males and less than 0.2 for females, with discrepancies potentially indicating sample mixups or low data quality); C) randomly selected one individual from each of the three pairs of individuals with duplicated genotyping for inclusion in the analysis.

Replication of GWAS signals using French sample

We assessed the evidence for replication of the two genome-wide significant SNPs in the independent French CF Modifier Gene Study. The sample quality control was aligned with the CGMS analysis. And we defined the sample inclusion criteria from an earlier candidate gene study in the French cohort (Mésinèle et al., 2022). Specifically, we restricted our replication analysis to the most recent two decades of available data (i.e., individuals born between 1991 and 2009). The above procedures resulted in an analysis sample of 501 individuals.

The phenotype used for analysis differs slightly between the Canadian and French analyses where the French study defines chronic *Pa* infection as at least three positive *Pa* cultures at least 1 month apart over a 6-month period, and the date of chronic *Pa* is that at the third isolation. The Canadian study defines chronic infection as the presence of three consecutive years of records with at least one positive culture in at least two of the three years, and the subject's age on the date of the first culture that met this criterion was entered as the age of chronic infection.

Sample relatedness

After performing QC, we used PLINK (version 2.00aLM; Web Resources) to identify any cryptic familial relationships among all individuals. We calculated KING-robust kinship estimator $\hat{\phi}$ [4] and applied the following conventional cutoff to infer sample relatedness:

monozygotic twins (MZ-twins) or duplicate samples if $\hat{\phi} < 0.354$; first-degree relations (full siblings and parent-offspring pairs) if $\hat{\phi} < 0.177$; second-degree relations if $\hat{\phi} < 0.0884$. We further calculated pairwise IBD sharing probabilities, Z_0, Z_1 and Z_2 (Figure i) to visualize and validate the inferred sample relatedness. Finally, for each parent-offspring or MZ-twin pair, we chose the ones included in previous studies or the ones with higher genotyping call rates if none were included in previous association analyses.

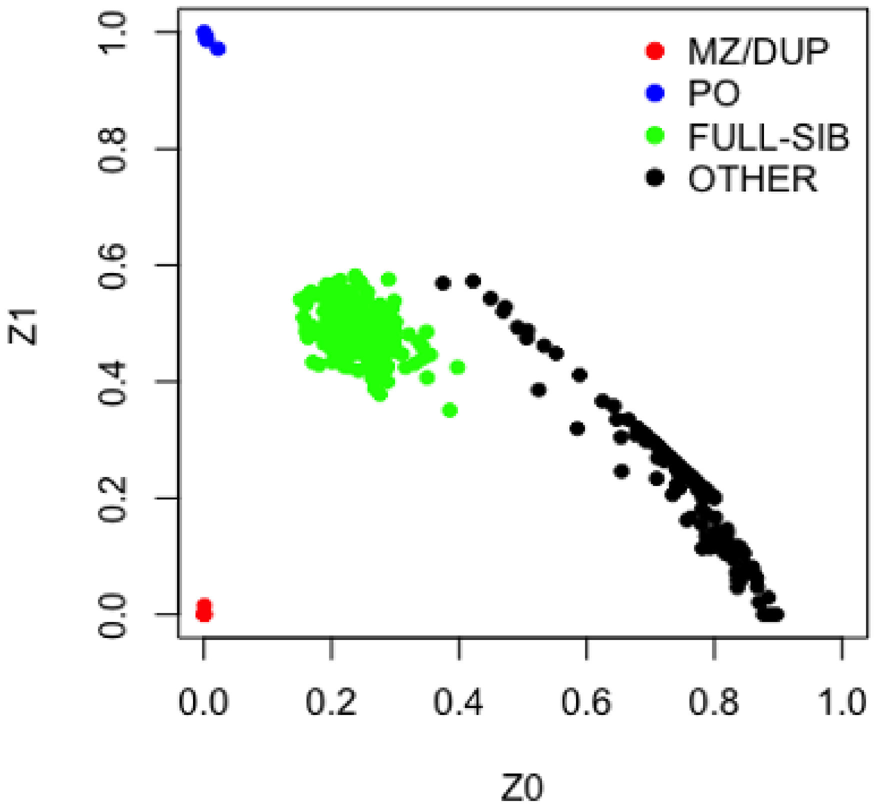


Figure i: Inferred sample relatedness by Identity by Descent (IBD) sharing estimated from KING [4]. Here Z_i is the estimated fraction sharing *i* copies between 2 individuals (*i* = 0, 1 or 2). The following criteria are used to infer sample relatedness: Unrelated individuals: (Z_0>0.3 & Z_2<0.1); Sibling pairs: (Z_0>0.1 & Z_1>0.3 & Z_2>0.1); Parent-offspring pairs: (Z_0<0.1 & Z_1>0.8 & Z_2<0.2); MZ-twins/duplicates: (Z_2>0.9).

Population structure

To infer and control population stratification, we calculated PCs by R [5] function ‘pcair()’ in the GENESIS package (version 2.12.4) [6]. This method was applied due to its advantage in accurately estimating population structure, even in the presence of familial relatedness. To calculate the PCs, we included only common SNPs with MAF>0.05 and in low linkage disequilibrium (LD) with each other (r^2<0.2).

To extract European samples for a discovery analysis, we anchored our study sample by the 1000 Genomes sample (2010/08/04 release, N=629) [7] whose ancestries are known. This was achieved by merging the overlapping markers of the genotyping data of our CGMS sample with the 1000 Genomes sample, and individuals were deemed of European ancestry if principal components 1 and 2 fell within a mean±6 s.d. rectangle formed using principal components from the 1000 Genomes datasets identified as European samples (Figure ii). The remaining samples were considered non-European samples.

Then for each GWAS sample, we run ‘pcair()’ and select the number of PCs for the downstream GWAS based on the Scree plots (Figure iii) of calculated eigen-values. Four PCs were included in the analysis of the European sample, in the non-European group and in the combined samples according to the ‘pcair()’ results.

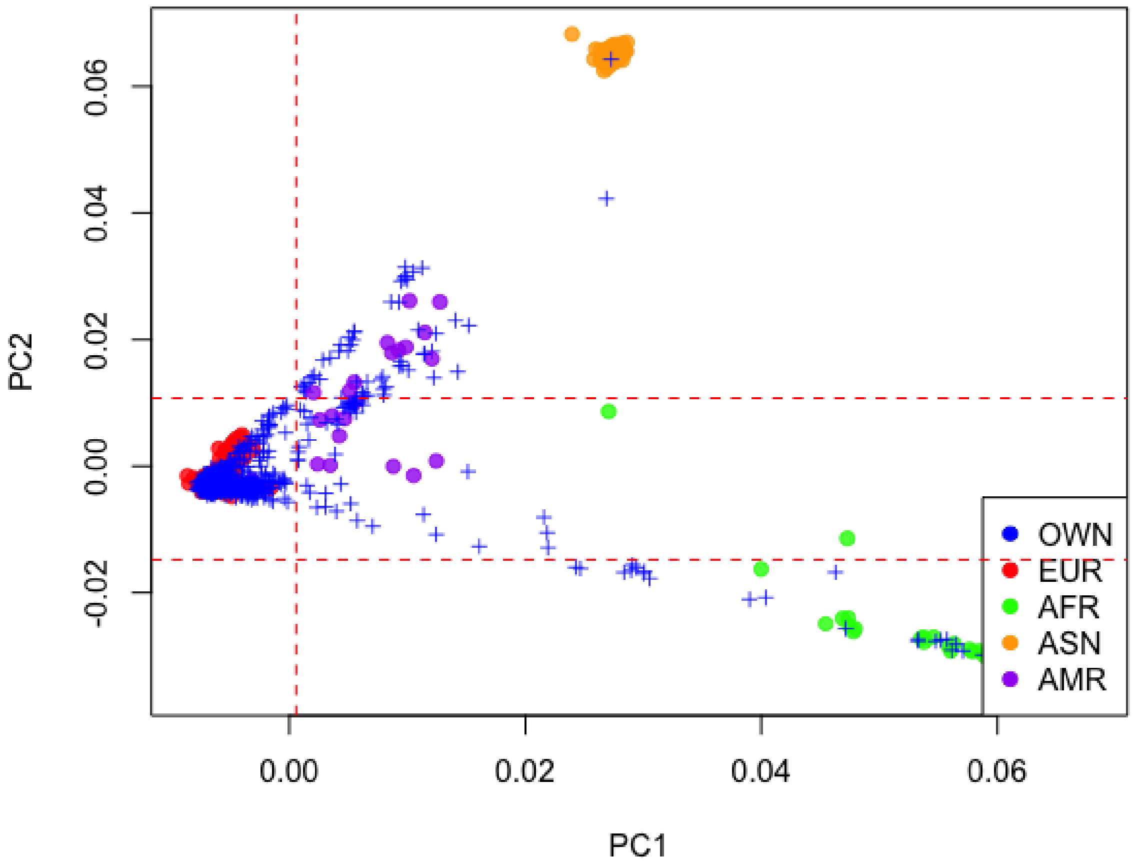


Figure ii: PC-Airs anchoring plot to identify sample ethnic groups. The PC-Airs are calculated based on common SNPs with MAF>0.05 and in low linkage disequilibrium (LD) with each other (r^2<0.2). Each circle represents a participant from the 1000 Genomes dataset (2010/08/04 release, N=629) colored by known ancestry, and each blue cross is a CGMS participant. CGMS participants were classified as of European ancestry if their scores on principal components 1 and 2 fell within a rectangle defined by the mean ± 6 standard deviations (s.d.) of the principal components 1 and 2 of the European samples from the 1000 Genomes Project (indicated by red dashed lines).

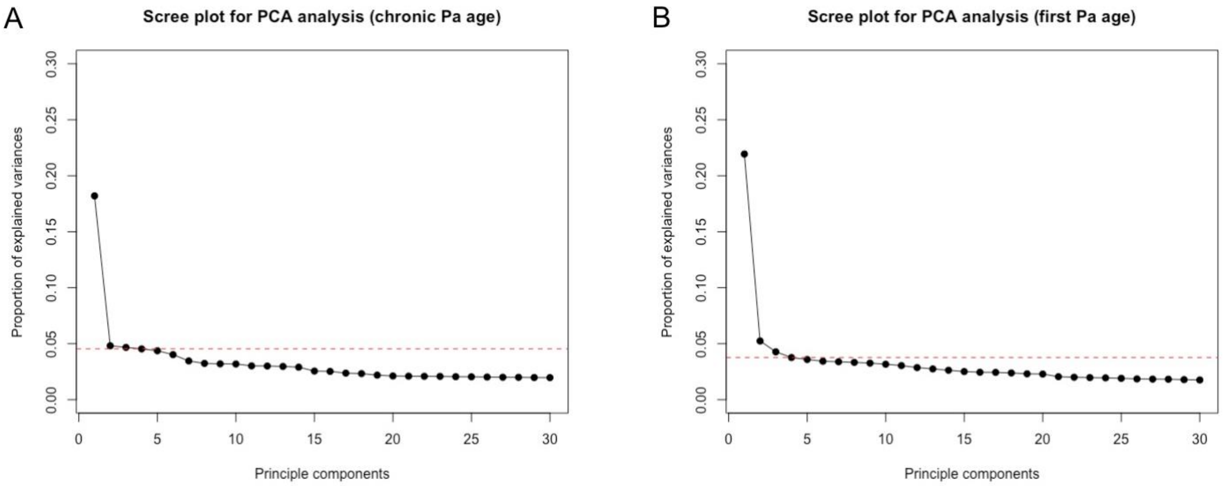


Figure iii: Scree plot for GWAS samples showing the proportion of variances explained by each of the top 30 principal components. The X-axis indicates the PC-Airs [6] calculated from (A) chronic *Pa* infection age and (B) first *Pa* age GWAS samples. Y-axis is the eigenvalues corresponding to each PC-Air, which quantifies the proportion of variances in genotype data explained by PC-Airs.

Imputation

Imputation was conducted in earlier studies of CF comorbidities [2, 8]. Briefly, genome-wide imputation of the full CGMS cohort was based on a hybrid reference panel integrating whole genome sequence from the 1000 Genomes Phase 3 data [7] and a subset (N=101) of the CGMS cohort. This subset reflected the demographic with severe cystic fibrosis in Canada and was subject to whole genome sequencing by Complete Genomics at an average coverage of 30X [8]. To ensure an accurate representation of *CFTR* disease-related haplotypes, which are poorly captured by standard public reference panels like the 1000 Genomes Phase 3, we utilized a bespoke reference that combines in-sample whole genome sequences with these critical variants [7].

Next, Beagle version 4.1 [9] was employed for the phasing and imputation of missing genotype data, which accounted for variations in genotyping platforms [8]. Single nucleotide polymorphisms (SNPs) that displayed an imputed Beagle quality score of r^2 lower than 0.8 were recorded as missing. Moreover, SNPs displaying multiple alternative alleles were omitted.

Residualized phenotypes

The distributions of the first Pa age and chronic Pa age in our sample are highly skewed to the right (Figure v (A) and (C)) and association tests on such traits can have inflated type I error or reduced power [10, 11]. To verify the residual normality assumption, we conducted linear regression with $Pa\ age = \alpha + \beta_{Age} Age + \epsilon$. Since the residuals are symmetric, we used the residuals as the response variables for association tests.

There are several advantages of applying this two-stage procedure and using the residualized phenotypes as the trait of interest. 1) The residuals can be interpreted as calendar year-adjusted infection ages. The interpretation is clearer than with directly or indirectly rank-based normal transformation outcomes. 2) It controls the potential confounding of the calendar effect. The current age is positively associated with both phenotypes (p<0.001), indicating a strong calendar trend of age-at-onset or diagnosis for *Pa* infection.

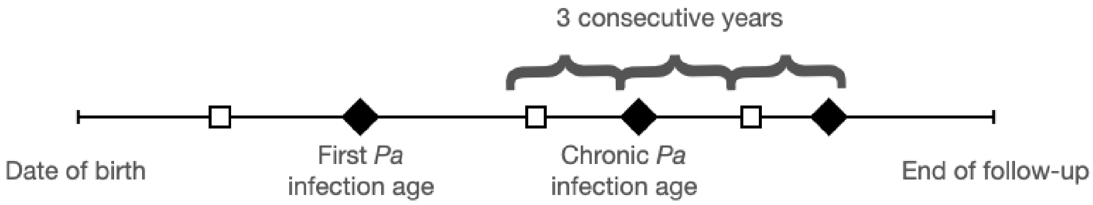


Figure iv: Illustration of the first and chronic *Pa* infection ages defined in [12] to capture the extremes in stages of respiratory infection. The first infection was defined as the first positive culture for *Pa*; the chronic infection was defined as the presence of 3 consecutive years of culture records with at least one positive culture in at least 2 of the 3 years. The subject's age on the date of the first culture that met this criterion was entered as the age of chronic infection. Subjects were defined as negative if no *Pa* had been cultured from the respiratory tract as of the end of follow-up. To ensure accurate assignment of the first positive culture, only those subjects with a previous negative culture for *Pa* before the first positive culture were included in the analysis.

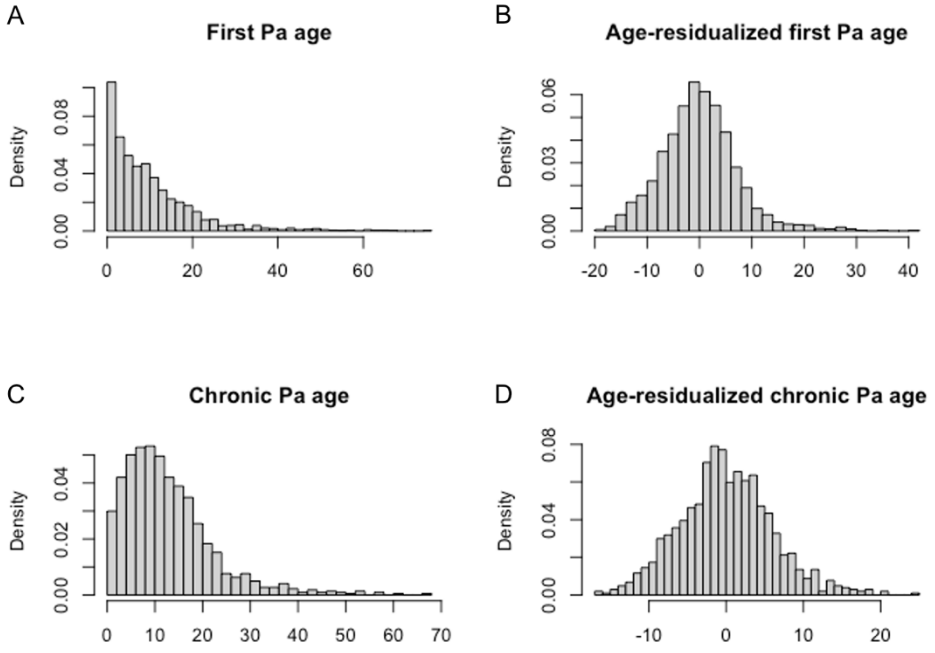


Figure v: The distributions of (A) the first *Pa* age and (C) the chronic *Pa* age in CGMS samples, and the corresponding residualized phenotypes (B) and (D). The distributions of the first *Pa* age and chronic *Pa* age are highly skewed to the right. The residuals, after regressing on the current age to adjust the calendar effect, appeared to be symmetric.

Linear mixed-effect model adjusting for sample relatedness

The association test is conducted under the following models:

$$\boldsymbol{y} = \boldsymbol{\alpha} \ + \boldsymbol{\beta_1} \boldsymbol{Age} + \boldsymbol{\epsilon}$$

$$\hat{\boldsymbol{\epsilon}} \ = \boldsymbol{\alpha}' + \boldsymbol{\beta_G} \boldsymbol{G} + \boldsymbol{\beta_S} \boldsymbol{sex} + \sum_{\mathbf{i}} \boldsymbol{PC_i} \ + \ \boldsymbol{g} + \boldsymbol{\epsilon}',$$

where $\boldsymbol{y}$ is the response variable, either the first *Pa* age or the chronic *Pa* age. $\hat{\boldsymbol{\epsilon}}$ is the residuals of the first stage and is used as the response variable at the second stage, and $\boldsymbol{g}$ is a random effect accounting for sample relatedness. $\boldsymbol{g} \sim \ \boldsymbol{N}(\mathbf{0}, \boldsymbol{\sigma_g} \boldsymbol{\Psi})$, $\boldsymbol{\Psi}$ is a matrix of pairwise measures of genetic relatedness, which are estimated by Kinship estimates from the genetic software KING.

Heritability estimation

We utilized Genome-wide Complex Trait Analysis (GCTA) [13] to estimate heritability, the fraction of phenotypic variance explained by genome-wide genetic variants. GCTA applies a restricted maximum likelihood (REML) method, assuming proportional genetic and phenotypic similarities among individuals to determine the variance explained by common genetic variants. In assessing heritability among unrelated individuals, GCTA exploits inherent genetic similarity stemming from shared ancestry to partition phenotypic variance into genetic and environmental components, using genotypically inferred relatedness. Essentially, if unrelated individuals share more common genetic variants than expected, they are more likely to show similar phenotypes, enabling GCTA to estimate trait heritability.

Table i. Heritability estimates of the first and chronic *Pa* infection age

Phenotype	**First *Pa* age**	**Chronic *Pa* age**
N	1,476	937
$\boldsymbol{h^2}$ **(SE)**	0.225 (0.175)	0.460 (0.280)
p-value	0.090	0.044

The first and chronic *Pa* age traits are residualized, after regressing the phenotype on the current age. The estimation was conducted using unrelated European samples from the Canadian CF Gene Modifier Study (CGMS).

X Chromosome analysis

For each X chromosome SNP, female genotypes were coded additively as 0, 1 and 2 and male genotypes were coded as 0 and 2, corresponding to the random X-inactivation assumption.

PheWAS analysis

PheWAS were performed using the GWAS atlas [14] (Web Resources).

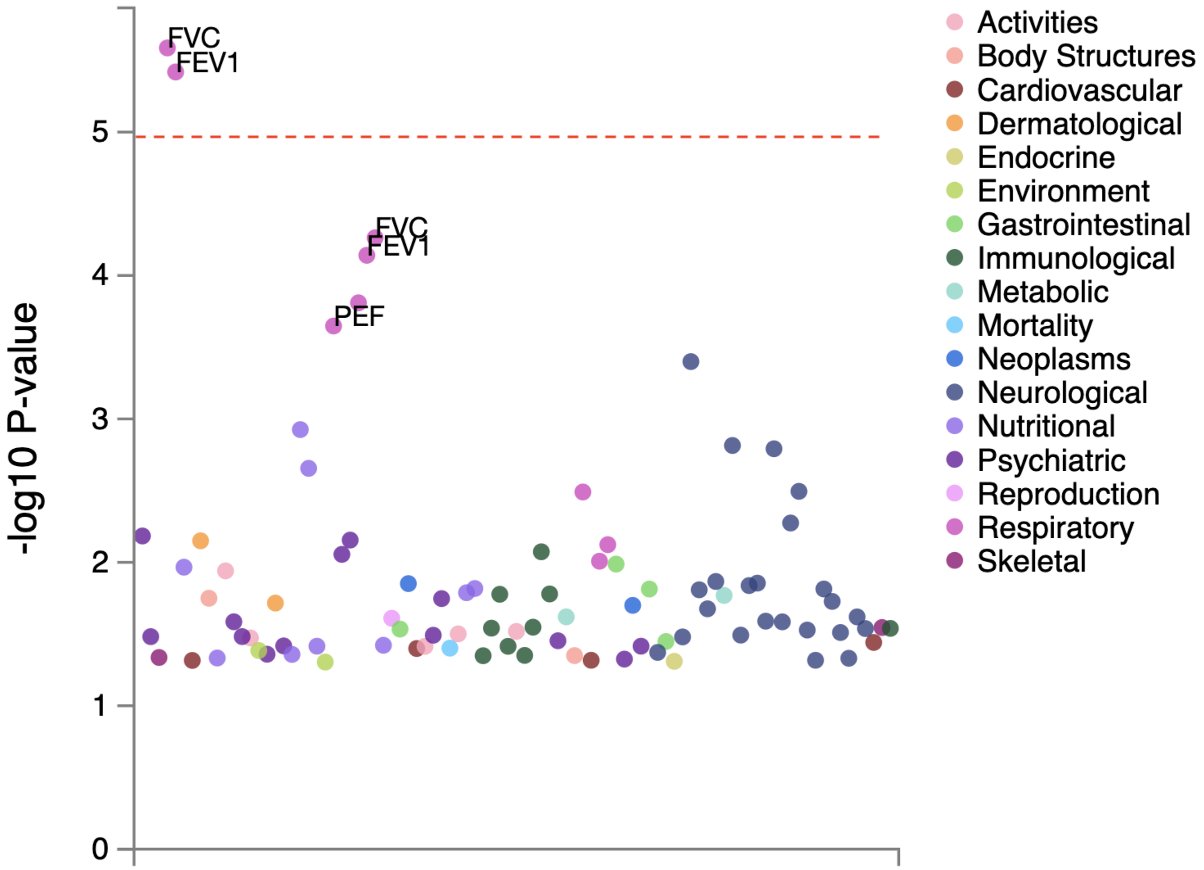


Figure vi: PheWAS results for rs62369766 (on chromosome 5) identified by chronic *Pa* age GWAS. Each dot represents a p-value (on the -log10 scale) for the association between rs6239766 and a trait from one of 4,756 GWAS ATLAS phenotypes [14]. For conciseness of visualization, only associations with p-values < 0.05 were plotted. Each dot is coloured based on the organ system of the trait [14]. The pair of points for each of FVC and FEV_1 corresponds to GWAS conducted on the UK Biobank and SpiroMeta Consortia [15], and the point for PEF corresponds to GWAS conducted on the UK Biobank. Since the associations of rs62369766 with FEV1/FVC were > 0.05 in Shrine et al. (2019; Table 3), they were not included. The red line is the $-\log_{10}(1.05 \times 10^{-5})$, corresponding to the conservative Bonferroni correction for considering the 4,756 tests at the nominal family-wise error rate of 0.05. The x-axis reflects the order of the sample sizes for GWAS for the traits (decreasing from left to right).

eQTL for genome-wide significant SNPs

We investigated evidence of eQTLs for the genome-wide significant SNPs at the two loci for chronic *Pa* using genotype tissue expression consortium (GTEx) data v8. We examined the expression of genes in relevant tissues including lung and whole blood. The nominal p-value for each variant-gene pair was obtained by testing whether the slope of a linear regression model between genotype and expression deviates from 0 [16].

Supplementary Appendix 2

Cross-trait polygenic risk score

Cross-trait polygenic risk score in individuals with CF

We conducted a cross-trait polygenic risk score (PRS) analysis to assess the genetic overlap between chronic *Pa* age and SaKnorm. The base or discovery data are the summary statistics from a CF lung function GWAS [1] and the target data are the genotype data from 120 unrelated European individuals from CGMS who were not included in the SaKnorm GWAS [1] to avoid data overfitting.

We determine the number of SNPs used to construct the PRS based on an optimal p-value threshold. To construct the PRS score, we used the following conventional Clumping + Thresholding (C+T) procedures.

For each *p* in the range of p-value thresholds (5E-8 to 1 in increments of 0.00005):

1. We built a corresponding PRS using SNPs with p-values below the cut-off. The PRS is constructed as the weighted sum of genotypes from the individuals included in the chronic *Pa* age GWAS, with the weights obtained from the summary statistics of the previously published CF lung function GWAS [1].
2. We then fit the regression model: *Pa* age ~ PRS + age + sex + top four PCs
3. We calculated the proportion of variation explained by PRS.

The optimal p-value threshold is then the value of *p* that results in the PRS score with the largest proportion of phenotypic variance explained in chronic *Pa* age residual. Therefore, the number of SNPs used in the PRS construction can differ between each of the outcome measures.

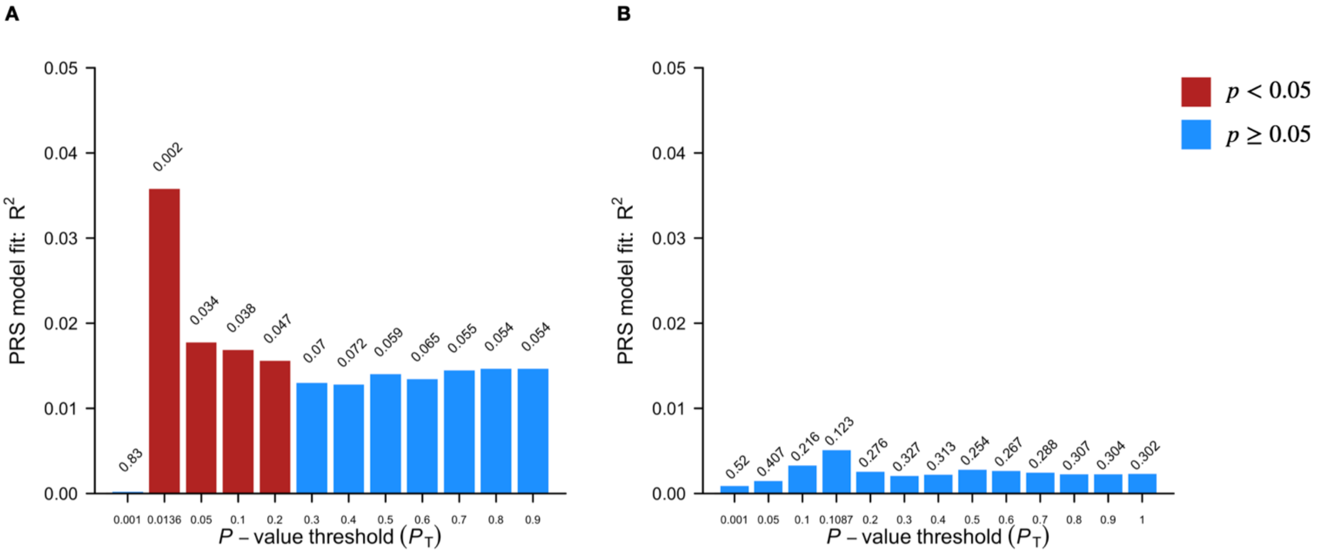


Figure i: Proportion of variation in chronic *Pa* age (A) and first *Pa* age (B) explained by the polygenic risk score (PRS) constructed from the summary statistics of the GWAS for SaKnorm in CF (Corvol et al., 2015). The x-axis presents a subset from a broad range of p-value thresholds (5E-8 to 1 in increments of 0.00005) at which the clumped SNPs from the lung function GWAS were selected to construct the PRS. The y-axis is the model fit corresponding to PRS, which is defined as the R^2 difference between the full regression model ($Pa\ age \sim PRS + age + sex + top\ four\ PCs$) and null model ($Pa\ age \sim sex + age + top\ four\ PCs$). The numbers above each bar represent the p-value threshold for PRS in the model. Red indicates that the corresponding PRS was significant at the 0.05 level, and blue indicates the corresponding PRS was not significant at the 0.05 level.

Table i: Top ten independent SNPs with the largest effect sizes on chronic *Pa* age in CF lung function PRS

SNP	CHR	BP	A1	A2	Mapped gene	Ensembl ID	Effect on CF lung function	p-value on CF lung function	Effect on chronic *Pa* age	p-value on chronic *Pa* age
rs58836625	5	132532778	G	A	*FSTL4*	ENSG00000053108	-0.181	9.10×10^{-4}	-2.89	8.15×10^{-3}
rs11572236	1	60383209	A	G	*CYP2J2*	ENSG00000134716	0.166	6.77×10^{-3}	-2.69	0.028
rs71579840	2	215674376	A	G	*BARD1*	ENSG00000138376	-0.140	7.06×10^{-3}	-2.58	0.015
rs7017918	8	51530583	G	A	*SNTG1*	ENSG00000147481	-0.144	3.94×10^{-3}	2.51	0.012
rs11759671	6	53007415	A	G	*GCM1*	ENSG00000137270	0.200	2.39×10^{-3}	2.48	2.77×10^{-3}
rs10977522	9	9137661	G	T	*PTPRD*	ENSG00000153707	0.159	0.012	-2.45	0.038
rs61784742	1	65380623	T	C	*JAK1*	ENSG00000162434	-0.143	2.90×10^{-3}	-2.35	0.012
rs11605831	11	11306150	A	G	*GALNT18*	ENSG00000110328	0.122	2.47×10^{-3}	-2.23	6.75×10^{-3}
rs112060594	17	34154197	A	G	*/*	ENSG00000270647	0.153	5.46×10^{-5}	2.19	2.50×10^{-4}
rs73099611	7	43202717	A	G	*HECW1*	ENSG00000002746	-0.135	4.80×10^{-3}	-2.19	0.020

Cross-trait polygenic risk score from a non-CF population

Additionally, we repeated the PRS analysis using GWAS summary statistics [15] (N=321,047) of forced vital capacity (FVC), forced expiratory volume in one second (FEV1), peak expiratory flow (PEF) and FEV_1/FVC using population-based cohorts from the UK Biobank (Bycroft et al., 2018). We followed the steps in the previous section to construct four PRS analyses (i.e., FEV_1/FVC-, FVC-, FEV1- and PEF-derived PRS) for chronic *Pa* infection age, and another set of PRS analyses for the first *Pa* infection age. For FEV_1/FVC, at the optimal p-value threshold at 5×10^{-5} (Table ii), the corresponding PRS constructed from 1,612 independent SNPs explained about 0.57% of the phenotypic variance of chronic *Pa* age (Figure ii (A)). Consistent with the CF-specific analysis, there was no significant association between the FEV_1/FVC-derived PRS and first *Pa* age (Figure ii (B)). We replicated the analysis using FVC, FEV_1 and PEF-derived summary statistics (Table ii) and had generally consistent results (Figures ii-v).

Table ii: Summary of PRS constructed from lung function phenotypes and their associations with chronic *Pa* age.

* Base data phenotype	Base data sample size	† Target data sample size	‡ Optimal p-value threshold	Number of SNPs included in PRS	Effect of PRS (year)	§ p-value of PRS	ll R^2 for PRS
SaKnorm	6,365	120	0.014	7,756	0.618	0.002	0.036
FEV$_1$/FVC	321,047	937	5×10^{-5}	1,612	-0.802	0.002	0.006
FVC			0.105	29,030	-0.241	0.168	0.001
FEV$_1$			0.041	17,984	-0.433	0.016	0.004
PEF			0.008	7,175	-0.643	0.002	0.006

* Base data phenotype is the phenotype on which we examine its genetic correlation with *Pa* age by using the summary statistics of corresponding GWAS as weights to construct PRS. We derived summary statistics for SaKnorm (the first row) from [1], the largest GWAS of CF lung disease to date from the International CF Gene Modifier consortium (N=6,365), and lung function phenotypes (row 2-4) from GWAS based on non-CF individuals from the UK Biobank European population [15].

† Target data are the genotypic data of the *Pa* age GWAS sample subset after removing the overlapped individuals included in the base data.

‡ Optimal p-value thresholds are selected from the sliding p-value threshold approach. At optimal p-value p_t, PRS constructed from SNPs with p-values less than p_t explained the largest proportion of phenotypic variance in chronic *Pa* age residual.

§ The p-values for PRS are calculated from (Pa age $\sim$ PRS $+$ sex $+$ age $+$ top four PCs).

ll R^2 for PRS is defined as the R^2 differences between the full regression model (Pa age $\sim$ PRS $+$ sex $+$ age $+$ top four PCs) and null model (Pa age $\sim$ sex $+$ age $+$ top four PCs).

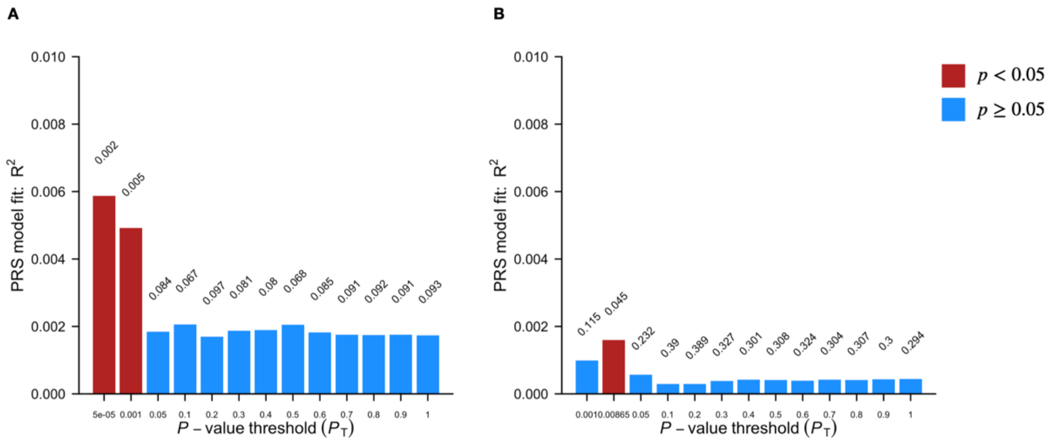


Figure ii: Proportion of variation in the chronic *Pa* age (A) and the first *Pa* age (B) explained by the PRS constructed from the GWAS summary statistics of FEV1/FVC ratio. The x-axis displays a subset of p-value thresholds ranging from 5E-8 to 1, in increments of 0.00005. These thresholds were used to select clumped SNPs from the lung function GWAS reported by [15] for constructing the PRS. The Y-axis is the model fit corresponding to PRS, which is defined as the R^2 differences between the full regression model (Pa age ~ PRS + age + sex + top four PCs) and null model (Pa age ~ sex + age + top four PCs). Red indicates that the corresponding PRS was significant at the 0.05 level, and blue indicates the corresponding PRS was not significant at the 0.05 level.

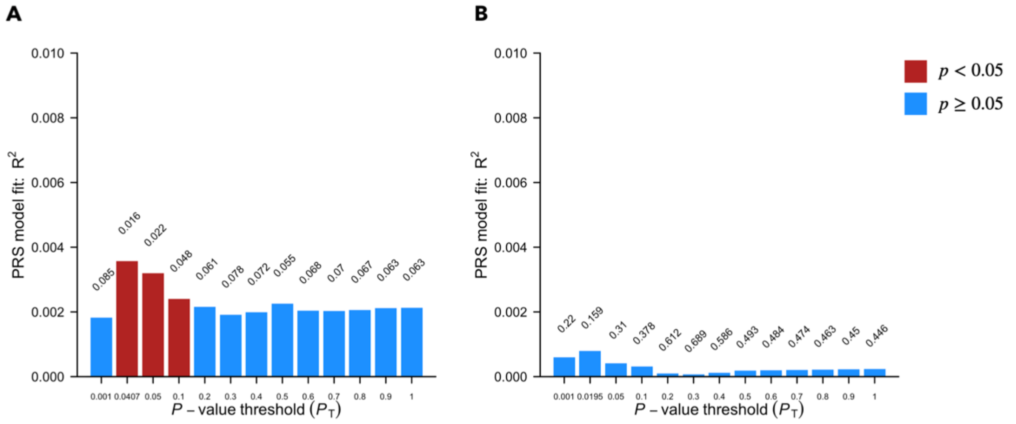


Figure iii: Proportion of variation in the chronic *Pa* age (A) and the first *Pa* age (B) explained by the PRS constructed from the GWAS summary statistics of FEV1. The x-axis displays a subset of p-value thresholds ranging from 5E-8 to 1, in increments of 0.00005. These thresholds were used to select clumped SNPs from the lung function GWAS reported by [15] for constructing the PRS. The Y-axis is the model fit corresponding to PRS, which is defined as the R^2 differences between the full regression model (Pa age ~ PRS + age + sex + top four PCs) and null model (Pa age ~ sex + age + top four PCs). Red indicates that the corresponding PRS was significant at the 0.05 level, and blue indicates the corresponding PRS was not significant at the 0.05 level.

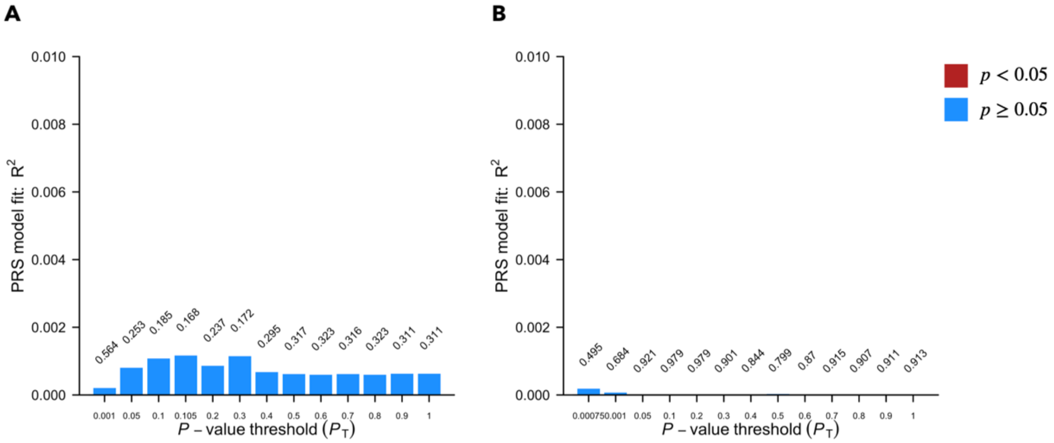


Figure iv: Proportion of variation in the chronic *Pa* age (Left) and the first *Pa* age (Right) explained by the PRS constructed from the GWAS summary statistics of FVC. The x-axis displays a subset of p-value thresholds ranging from 5E-8 to 1, in increments of 0.00005. These thresholds were used to select clumped SNPs from the lung function GWAS reported by [15] for constructing the PRS. The Y-axis is the model fit corresponding to PRS, which is defined as the R^2 differences between the full regression model (Pa age $\sim$ PRS + age + sex + top four PCs) and null model (Pa age $\sim$ sex + age + top four PCs). Red indicates that the corresponding PRS was significant at the 0.05 level, and blue indicates the corresponding PRS was not significant at the 0.05 level.

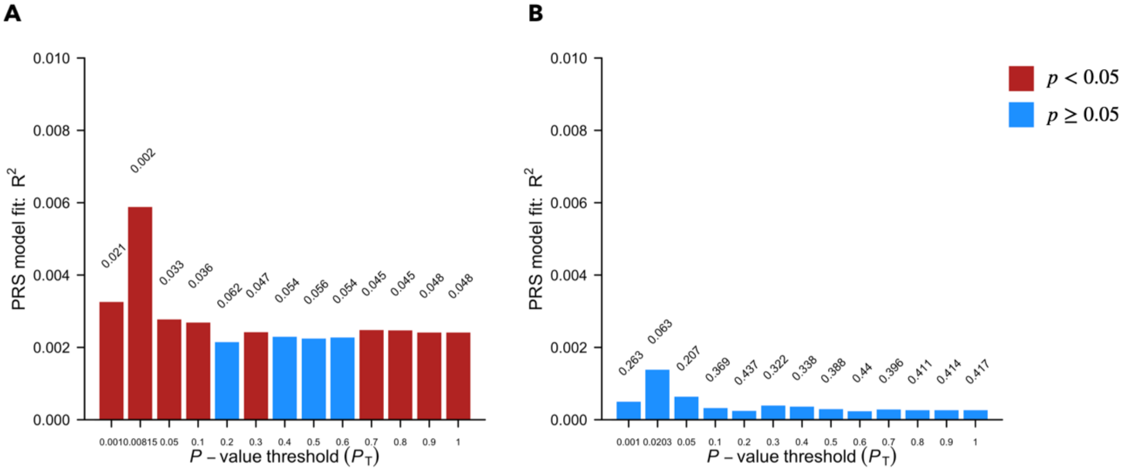


Figure v: Proportion of variation in the chronic *Pa* age (Left) and the first *Pa* age (Right) explained by the PRS constructed from the GWAS summary statistics of PEF. The x-axis displays a subset of p-value thresholds ranging from 5E-8 to 1, in increments of 0.00005. These thresholds were used to select clumped SNPs from the lung function GWAS reported by [15] for constructing the PRS. The Y-axis is the model fit corresponding to PRS, which is defined as the R^2 differences between the full regression model (Pa age $\sim$ PRS + age + sex + top four PCs) and null model (Pa age $\sim$ sex + age + top four PCs). Red indicates that the corresponding PRS was significant at the 0.05 level, and blue indicates the corresponding PRS was not significant at the 0.05 level.

Supplementary Appendix 3

Bi-directional Mendelian Randomization

We used the TwoSampleMR package [17] to conduct bidirectional Mendelian Randomization (MR) analysis to investigate causal relationships between chronic *Pa* infection and lung disease in CF. We test whether the onset of chronic *Pa* infection (measured by chronic infection age) affects lung function (measured by SaKnorm), and vice versa.

<u>Selection of instrumental variables (IVs)</u>

To enhance statistical power, we implemented a multivariate IVs analysis, selecting SNPs to serve as IVs based on a polygenic risk score (PRS) analysis (akin to the cross-trait PRS analysis but applied to a single trait here). First, we randomly split the CGMS GWAS samples (N=1,037) into a base sample (N=520) and a target sample (N=517). We then conducted a GWAS on the base sample for chronic *Pa* age, using the set of SNPs pruned for linkage disequilibrium (LD r^2 < 0.1). From the GWAS summary statistics, we constructed a series of PRSs for the target sample corresponding to a range of p-value thresholds (5E-8 to 0.5 in increments of 0.00005) implemented through PRSice-2 [18]. For each threshold, we built the PRS using SNPs with p-values below the cut-off and quantified the explained variance in chronic *Pa* age via regression R^2. The PRS constructed with p-value < 0.011 yielded the highest R^2. Following further LD clumping of SNPs within this PRS (r^2<0.001; as MR is more sensitive to independence of IVs [17]), we identified N=677 independent SNPs to serve as IVs for chronic *Pa* age.

Reversely, we tested whether reduced lung function leads to earlier chronic *Pa* age following the same steps. We used the SNPs included in cross-trait PRS constructed from summary statistics of CF lung function GWAS [1] with an optimal p-value threshold of 0.014. After further LD clumping (r^2<0.001), we used N=771 independent SNPs as IVs for SaKnorm.

<u>IV assumptions</u>

We first evaluate the three key assumptions of MR:

1. Relevance: We evaluate the relevance of IVs by the F-statistic from the regression of exposure on PRS constructed by IVs and compare it with the rule-of-thumb threshold of 10. The F-statistic is 9.38 (calculated by the formula $F = \frac{R^2}{1-R^2}\frac{n-k}{k-1}$ where n is the sample size, and k is the number of predictors) for PRS of chronic *Pa* age, indicating that the IVs are moderately strong, but the MR result should be interpreted with the caution of weak instrument bias. On the other hand, the F-statistic is 16.01 for PRS of SaKnorm, indicating that IVs are robustly associated with SaKnorm.

2 No horizontal pleiotropy: We evaluated horizontal pleiotropy by testing whether the intercept term in the MR-Egger regression significantly departures from zero [19].

3. Exchangeability: The Exchangeability assumption posits that the IVs are not correlated with any confounding factors that could introduce bias in the association between the exposure and the outcome. In practical scenarios, directly evaluating this assumption can be challenging. Nevertheless, we proceed with this assumption because there is no evidence of population stratification, dynastic effects, or assortative mating, which are potential sources of violation of the independence assumption [20, 21].

Since we included hundreds of SNPs as IVs, we further tested the homogeneity in causal effect estimation by Cochran's Q test.

MR analysis

The causal effect of exposure on outcome was estimated by the inverse variance weighted (IVW) method [22] implemented in the TwoSampleMR package [17].

Sensitivity analysis

Sensitivity analysis (Online Methods) was conducted based on (1) various MR-methods accounting for bias arising from weak IVs (MR weighted median [23]), outliers (MR weighted median and Simple mode [24]) and horizontal pleiotropy (MR-Egger [19]); (2) p-value thresholds at 5E-8, 5E-7, 5E-6, 5E-5, 5E-4, 5E-3, 5E-2 for selecting IVs.

Table i. Mendelian Randomization analysis and sensitivity analysis

A. Causal effect of chronic *Pa* age on lung function

p-value threshold for IV	# of SNPs as IV	p-value					p-value for IV assumptions	
		IVW	MR-Egger	Weighted median	Simple mode	Weighted mode	Heterogeneity	Pleiotropy
0.011 (PRS)	677	0.01	0.75	0.38	0.76	0.98	0.09	0.46
5E-8	2	0.27	N/A	N/A	N/A	N/A	N/A	N/A
5E-7	2	0.27	N/A	N/A	N/A	N/A	N/A	N/A
5E-6	9	0.28	0.46	0.35	0.33	0.37	0.83	0.29
5E-5	40	0.91	0.75	0.74	0.45	0.39	0.27	0.77
5E-4	225	0.13	0.79	0.44	0.64	0.83	0.53	0.85
5E-3	625	0.03	0.59	0.31	0.80	0.92	0.64	0.20
5E-2	1138	2.3E-3	0.53	0.09	0.04	0.97	0.29	0.65

B. Causal effect of lung function on chronic *Pa* age

p-value threshold for IV	# of SNPs as IV	p-value					p-value for IV assumptions	
		IVW	MR-Egger	Weighted median	Simple mode	Weighted mode	Heterogeneity	Pleiotropy
0.014 (PRS)	527	4.2E-4	0.06	0.02	1.24	0.10	0.47	0.51
5E-8	4	0.54	0.90	0.78	0.69	0.79	0.02	0.83
5E-7	5	0.46	0.92	0.63	0.92	0.90	0.05	0.83
5E-6	15	0.39	0.46	0.57	0.65	0.73	0.07	0.58
5E-5	65	0.02	0.06	0.02	0.15	0.16	0.64	0.18
5E-4	225	3.4E-3	0.02	0.05	0.23	0.20	0.74	0.16
5E-3	479	7.6E-4	0.08	0.01	0.12	0.13	0.58	0.53
5E-2	762	1.0E-3	0.24	5.5E-3	0.09	0.13	0.55	0.97

Supplementary Appendix 4
GWAS of first *Pa* age

GWAS of first *Pa* age with European-descent individuals (n = 1,574 including 100 sibling clusters) identifies a genome-wide significant SNP rs6847677 (chr4p13; p-value=2.37×10^{-8};Intron variant on *EVC2*) and a suggestive locus with leading SNP rs60110289 (chrXp13; p-value=9.27×10^{-8}; Figure i). With the inclusion of non-European-descent individuals and sibling pairs (in total n=1,653 for first *Pa* age), however, the p-values increase to 4.73×10^{-8} and 5.41×10^{-7} respectively.

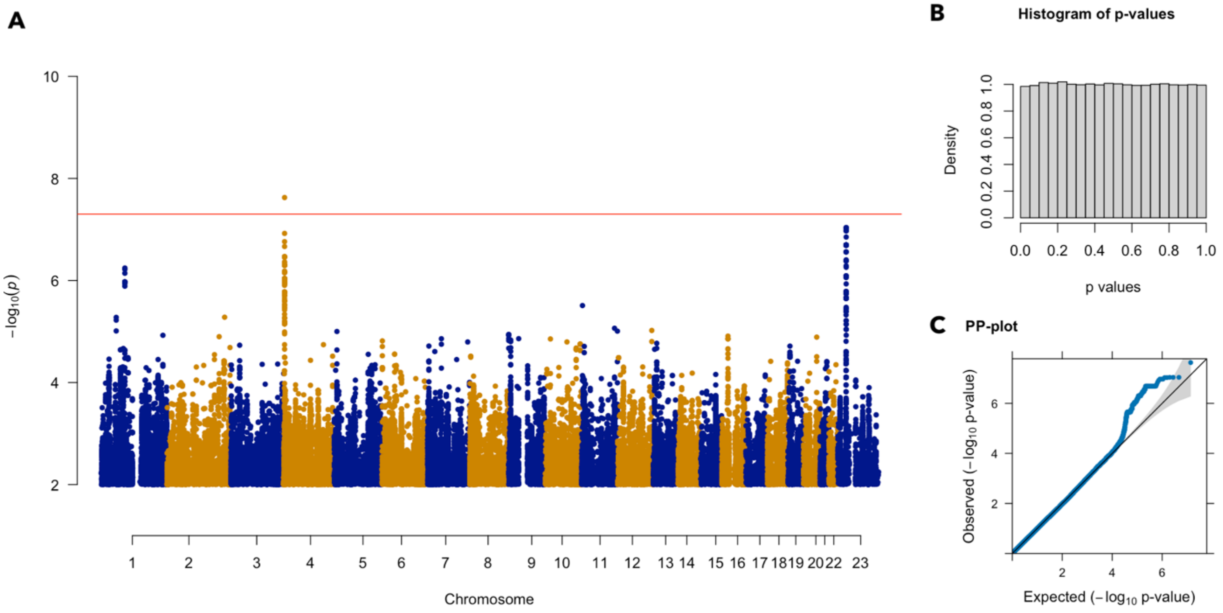


Figure i: The Manhattan (A), histogram (B) and Q-Q plot (C) of the p-values from GWAS of the first *Pa* age in N=1,574 European individuals. In total 6,994,213 SNPs were analyzed, all with MAF>5%. The red horizontal line corresponds to the genome-wide significance threshold of p-value=5×10^{-8}.

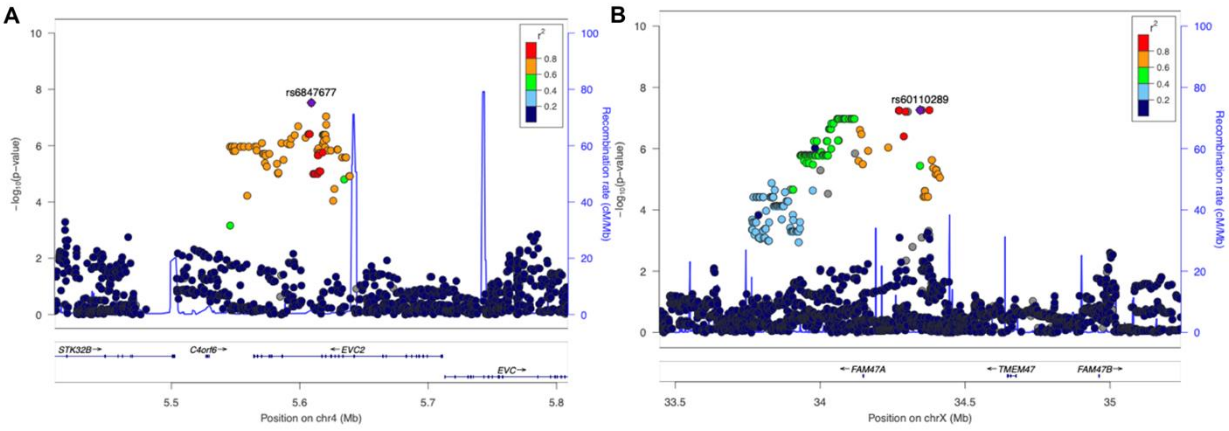


Figure ii: LocusZoom plots using the hg19 genome build for the two loci associated with the first *Pa* age. **(A)** rs6847677 on chromosome 4, **(B)** rs60110289 on chromosome X. Within each plot, the colors represent the 1000 Genomes EUR linkage disequilibrium r^2 values with the top SNP (shown as the purple diamond and labelled with dbSNP ID).

Supplementary Appendix 5
Sensitivity analysis on confounding factors

Our GWAS samples comprised 1,037 individuals, born from 1941 to 2010 (Supplementary Figures 3 and 4), and recruited from various Canadian CF clinics. To determine whether environmental factors such as CF diagnostic age, *CFTR* genotypes, birth cohorts, and recruitment sites confounded the SNP signals, we conducted comprehensive stratified sensitivity analyses. Furthermore, a survival analysis was implemented to adjust for possible survival bias.

1. The identified genetic associations are stronger in individuals who were diagnosed with CF at a younger age

We conducted a sensitivity analysis to see if the two GWAS signals are robust to the exclusion of individuals who were diagnosed with CF at an older age and therefore had a less severe course of disease. We found that both loci exhibited a consistent effect direction and significance with our primary GWAS (Table i). Moreover, signals at one locus (rs927553/rs927552) were more pronounced with a new genome-wide significant SNP at this locus identified, rs881428, when restricting the analysis to individuals diagnosed with CF by the age of 3.

Table i. Stratified analysis for CF diagnosis age

SNP	CHR	Effect Allele	Reference Allele	Sample stratified by CF diagnostic age	MAF	Beta (Year)	p-value
rs927553	13	G	C	All samples (N=1,037)	0.44	1.51	1.91×10^{-8}
				CF diagnostic age ≤3 (N = 840)	0.42	1.70	6.89×10^{-9}
				CF diagnostic age > 3 (N = 197)	0.49	0.66	0.32
rs927552	13	G	A	All samples (N=1,037)	0.41	1.51	3.64×10^{-8}
				CF diagnostic age ≤ 3 (N = 840)	0.39	1.72	8.49×10^{-9}
				CF diagnostic age > 3 (N = 197)	0.45	0.60	0.39
*rs881428	13	A	G	All samples (N=1,037)	0.16	1.38	1.81×10^{-7}
				CF diagnostic age ≤ 3 (N = 840)	0.17	1.61	2.20×10^{-8}
				CF diagnostic age > 3 (N = 197)	0.15	0.23	0.73
rs62369766	5	T	G	All samples (N=1,037)	0.16	-2.00	1.98×10^{-8}
				CF diagnostic age ≤ 3 (N = 840)	0.17	-1.69	8.87×10^{-6}
				CF diagnostic age > 3 (N = 197)	0.15	-3.39	2.70×10^{-4}

The genetic effects and p-values for the GWAS and stratified analysis are presented separately for three groups: the full cohort of 1,037 individuals regardless of age at CF diagnosis, a subgroup of 840 individuals diagnosed with CF at age 3 or younger, and 197 individuals diagnosed with CF at an age older than 3.

* An additional genome-wide significant SNP was identified from stratified analysis with the subgroup of 840 individuals diagnosed with CF at age 3 or younger. rs881428 is in strong LD with rs927553 ($r^2 = 0.97$).

2. The identified associations are not sensitive to *CFTR* genotypes

We repeated the GWAS for individuals homozygous for p.Phe508del and separately for those with a compound heterozygous genotype with p.Phe508del and a Class 1 mutation. Among the 1,037 individuals, 611 individuals are homozygous for p.Phe508del and 219 have compound heterozygous genotypes with a Class 1 mutation on their second allele. The effect sizes and direction of the three genome-wide significant SNPs, as inferred from the stratified GWAS, remain consistent with the GWAS of the entire sample. This suggests that the associated loci do not show evidence of being *CFTR* genotype-dependent.

Table ii. Stratified analysis for *CFTR* genotypes

SNP	CHR	Effect Allele	Reference Allele	Sample stratified by CFTR genotypes	MAF	Beta (Year)	p-value
rs927553	13	G	C	All samples (N=1,037)	0.44	1.51	1.91×10^{-8}
				delF508 homozygotes (N=611)	0.43	1.53	1.88×10^{-5}
				Compound heterozygotes (N=219)	0.42	1.85	1.40×10^{-3}
rs927552	13	G	A	All samples (N=1,037)	0.41	1.51	3.64×10^{-8}
				delF508 homozygotes (N=611)	0.39	1.56	1.79×10^{-5}
				Compound heterozygotes (N=219)	0.37	1.87	1.47×10^{-3}
rs62369766	5	T	G	All samples (N=1,037)	0.16	-2.00	1.98×10^{-8}
				delF508 homozygotes (N=611)	0.16	-1.99	2.39×10^{-5}
				Compound heterozygotes (N=219)	0.16	-1.75	2.14×10^{-2}

The genetic effects and p-values for the GWAS and stratified analysis are presented separately for three groups: the full cohort of 1,037 individuals irrespective of *CFTR* genotype, a subgroup of 611 individuals homozygous for p.Phe508del, and 219 individuals who have compound heterozygous genotypes with p.Phe508del and a Class 1 mutation on the second allele.

3. Stratified analysis based on birth cohort

The shift in the age of chronic *Pseudomonas aeruginosa* (*Pa*) infection onset from childhood to adulthood among Canadian cystic fibrosis (CF) patients over the past thirty years can be attributed to enhancements in treatments and healthcare methods. Before the 1980s, the focus of treatments was mostly on symptom management. Significant changes since then include an emphasis on rigorous nutritional strategies, a deeper understanding of CF's biological and genetic aspects, the implementation of proactive antimicrobial therapies, and the use of inhaled antibiotics (Hodson et al., 1981; Ramsey et al., 1993; Ratjen et al., 2010). To mitigate potential confounding effects from these treatments, the study population was divided into three birth cohorts, each representing the predominant CF treatments likely to have influenced them:

(1) Before 1980-12-31, N = 354
(2) Between 1981-01-01 and 1990-12-31 (N = 374), during which individuals with CF benefited from aggressive nutritional rehabilitation
(3) After 1991-01-01 (N = 309), at which point there has been consistent use of inhaled antibiotics

Our study demonstrates that the two GWAS signals consistently retained their statistical significance (at the 0.05 level) across three different birth cohorts. We observed a declining genetic impact from the oldest to the youngest cohort, possibly signalling that the improved treatment protocols have reduced the effect size of the SNP in more recent generations.

Noticeably, we observed a consistently strong positive age effect in combined analysis and stratified analysis in each cohort (Table iv), indicating that younger individuals tend to have earlier chronic infection ages. This strong age effect may potentially be driven by two factors: (1). Survival bias (see details and adjustment for the bias in the following Appendix 6*: Sensitivity analysis with time-to-event modelling*); (2). Treatment and changes in clinical care: the younger cohort might have greater access to more sensitive and frequent *Pa* diagnoses.

Table iii. Stratified analysis for birth cohorts

SNP	CHR	Effect Allele	Reference Allele	Sample stratified by Birth-cohort	N	MAF	Beta (Year)	p-value
rs927553	13	G	C	All samples	1,037	0.44	1.51	1.91×10^{-8}
				Prior to 1980-12-31	354	0.42	1.85	2.04×10^{-3}
				1981-01-01 ~ 1990-12-31	374	0.43	1.62	2.32×10^{-5}
				After 1991-01-01	309	0.46	0.97	3.50×10^{-3}
rs927552	13	G	A	All samples	1,037	0.41	1.51	3.64×10^{-8}
				Prior to 1980-12-31	354	0.39	2.28	1.95×10^{-4}
				1981-01-01 ~1990-12-31	374	0.38	1.48	1.72×10^{-4}
				After 1991-01-01	309	0.43	0.78	0.019
rs62369766	5	T	G	All samples	1,037	0.16	-2.00	1.98×10^{-8}
				Prior to 1980-12-31	354	0.18	-2.61	6.18×10^{-4}
				1981-01-01 ~1990-12-31	374	0.16	-1.78	4.47×10^{-4}
				After 1991-01-01	309	0.15	-1.21	0.010

The genetic effects and p-values for the GWAS and stratified analysis are presented separately for three groups: the full cohort of 1,037 individuals irrespective of birth cohort, a subgroup of 354 individuals who were born before 1980-12-31, 374 individuals who were born between 1981-01-01 and 1990-12-31, and 309 individuals born after 1991-01-01.

Table iv. Effect of age covariate in stratified analysis

Sample stratified by Birth-cohort	N	Beta of Age (Year)	95% C.I.
All samples	1,037	0.46	[0.43, 0.49]
Prior to 1980-12-31	354	0.68	[0.57, 0.80]
1981-01-01 ~ 1990-12-31	374	0.50	[0.32, 0.68]
After 1991-01-01	309	0.35	[0.27, 0.44]

The effects of age and 95% confidence intervals for the GWAS and stratified analysis are presented separately for three groups: the full cohort of 1,037 individuals irrespective of birth cohort, a subgroup of 354 individuals who were born before 1980-12-31, 374 individuals who were born between 1981-01-01 and 1990-12-31, and 309 individuals born after 1991-01-01.

Table v. Sensitivity analysis on CF individuals who did not undergo newborn screening (NBS).

SNP	CHR	Effect Allele	Reference Allele	Samples	MAF	Beta (Year)	p-value
rs927553	13	G	C	All samples (N=1,037)	0.44	1.51	1.91×10^{-8}
				CF individuals without NBS (N=1,020) *	0.43	1.51	3.64×10^{-8}
rs62369766	5	T	G	All samples (N=1,037)	0.16	-2.00	1.98×10^{-8}
				CF individuals without NBS (N=1,020) *	0.17	-2.01	2.33×10^{-8}

The genetic effects and p-values are presented separately for the GWAS of all 1,037 individuals and for the subset of 1,020 individuals not subjected to newborn screening (NBS). For sensitivity analysis, 14 individuals identified through NBS and 3 individuals with unknown NBS status were excluded.

Supplementary Appendix 6
Sensitivity analysis with time-to-event modelling

Our primary GWAS focused on individuals who had developed chronic *Pa* infection before the end of the follow-up. As a result, we must exercise caution when generalizing our findings to the entire Canadian CF population, since the GWAS sample may not be fully representative of this broader population.

Three sets of individuals with CF may have been excluded from our analysis:

1. Individuals who were never chronically infected with *Pa* before the end of the follow-up. (N=536)
2. Individuals who were chronically infected with *Pa* before the end of the follow-up, but the date for defining chronic *Pa* was missing due to gaps between longitudinal visits. (N=834)
3. Individuals who passed away before the start date of CGMS recruitment (most were recruited during 2002-2006, as shown in Supplementary Figure 3).

If a causal SNP is associated with both chronic *Pa* and survival/inclusion probability, there may be potential survival bias.

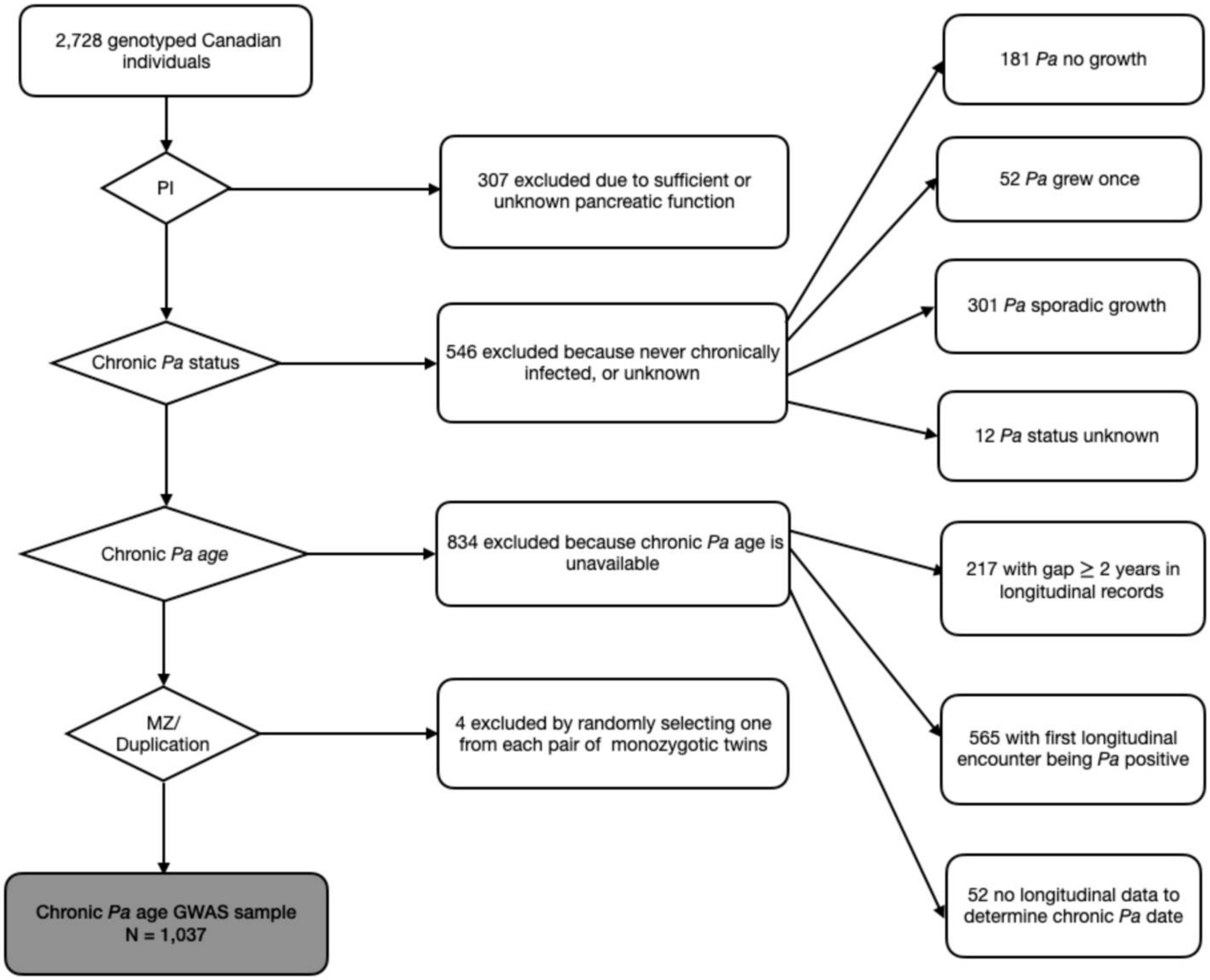

2,728 genotyped Canadian individuals
PI
307 excluded due to sufficient or unknown pancreatic function
Chronic *Pa* status
546 excluded because never chronically infected, or unknown
181 *Pa* no growth
52 *Pa* grew once
301 *Pa* sporadic growth
12 *Pa* status unknown
Chronic *Pa age*
834 excluded because chronic *Pa* age is unavailable
217 with gap ≥ 2 years in longitudinal records
565 with first longitudinal encounter being *Pa* positive
52 no longitudinal data to determine chronic *Pa* date
MZ/ Duplication
4 excluded by randomly selecting one from each pair of monozygotic twins
Chronic *Pa* age GWAS sample N = 1,037

Figure i: Sample distribution of the Canadian Cystic Fibrosis Gene Modifier Study (CGMS) cohort based on *Pseudomonas aeruginosa* (*Pa*) infection status as of December 31, 2020. From the cohort of 2,728 participants, we first excluded 307 individuals with sufficient or unknown pancreatic function. Next, we excluded 546 individuals who were not chronically infected (no growth, single growth, sporadic growth, or *Pa* status unknown), and 834 participants had chronic *Pa* infections with unspecified onset dates, attributed to: (a) missing longitudinal data exceeding two years before chronic infection confirmation (N=217), (b) lack of initial *Pa*-negative record (N=565), and (c) no or insufficient longitudinal data to establish the chronic infection timeline (N=52). Finally, we randomly excluded 1 individual from of the four pairs of monozygotic twins. The remaining 1,037 participants were analyzed in the GWAS.

For the first and second sets of individuals, we carried out time-to-event modelling that incorporated them with censored event time at their last longitudinal visits. Among the 2,728 genotyped individuals with insufficient pancreatic function, 534 did not exhibit chronic growth of *Pa* throughout the follow-up period, instead the *Pa* either never grew, grew only once, or sporadically grew (Figure i). Among these individuals 487 have the last longitudinal encounter time available (47 have missed longitudinal encounter time due to incomplete registry data) which were used to derived right-censored chronic infection-free event time. Similarly, 834 individuals were known to be chronically infected from the chart review record, but the date for defining chronic *Pa* was missing due to (a) gaps exceeding two years in longitudinal data before chronic infection confirmation (N=217), (b) lack of an initial *Pa*-negative record (N=565), and (c) no or insufficient longitudinal data to establish the chronic infection timeline (N=52). Out of which 792 have the last longitudinal encounter time available which were used to derive left-censored chronic infection-free event time.

Table i. Sensitivity analysis for time-to-event modelling

SNP	CHR	Effect Allele	Reference Allele	Sample (size)	MAF	Beta [95% C.I.]	exp(Beta) §	p-value
rs927553	13	G	C	Observed event (N=1,037)	0.44	1.51 [0.98, 2.04] (linear model*)	/	1.91×10^{-8}
						-0.25 [-0.34, -0.16]	0.78	8.36×10^{-8}
				Incl. left-censoring (N=1,829)	0.44	-0.25 [-0.34, -0.15]	0.78	3.46×10^{-7}
				Incl. right-censoring (N=1,524)	0.44	-0.14 [-0.24, -0.05]	0.87	1.90×10^{-3}
				Incl. left- and right-censoring (N=2,317)	0.44	-0.12 [-0.20, -0.05]	0.89	1.07×10^{-3}
rs62369766	5	T	G	Observed event (N=1,037)	0.17	-2.00 [-2.70, -1.30] (linear model*)	/	1.98×10^{-8}
						0.36 [0.23, 0.49]	1.43	1.40×10^{-7}
				Incl. left-censoring (N=1,829)	0.16	0.35 [0.22, 0.49]	1.42	2.64×10^{-7}
				Incl. right-censoring (N=1,524)	0.16	0.19 [0.06, 0.32]	1.21	3.36×10^{-3}
				Incl. left- and right-censoring (N=2,317)	0.16	0.16 [0.06, 0.26]	1.17	2.33×10^{-3}

Genome-wide significant SNPs' effects and p-values from GWAS and sensitivity analysis using a Cox proportional hazard model with censored data.

* Beta values for the GWAS were estimated using a linear mixed-effects model for samples with observed event times, to quantify the genetic impact of the coding allele on chronic *Pa* age. A positive Beta indicates that each additional copy of the allele is associated with an increased chronic *Pa* age, suggesting a potential protective effect. Conversely, Beta values from the Cox regression model, which represent the log-hazard ratio for each allele copy, imply that a positive value is associated with an increased rate of events. Hence, a protective effect is suggested by a positive Beta in the linear mixed-effects model and a negative Beta in the Cox model.

§ exp(Beta) estimated from the Cox-regression model quantifies the hazard ratio for an extra copy of the effect allele.

We performed a sensitivity analysis to evaluate the robustness of the genetic association with chronic *Pa* infection when incorporating individuals missing chronic infection age (Table i). First, by incorporating 792 individuals who were chronically infected with *Pa* but the exact dates of infection were missed, the hazard ratio for chronic *Pa* infection in individuals with an extra copy of the effect allele of rs927553 was estimated to be 0.78 (p-value=3.46×10^{-7}), consistent with the estimates based on GWAS samples. By incorporating 487 individuals who did not exhibit consistent growth of *Pa*, the hazard ratio was 0.87 with attenuated p-value at 1.90×10^{-3}, although the conclusion is consistent with the GWAS conclusion that each extra copy of the effect allele is associated with 1.51 years later onset of chronic *Pa* infection. Lastly, by incorporating both left- and right-censored samples, the hazard ratio was 0.89 (p-value=1.07×10^{-3}). A similar pattern was observed for rs62369766. In conclusion, while the estimates of effect are not influenced by the left-censored samples, they do show sensitivity to the right-censored samples. However, these sensitivities do not alter the overarching conclusions and interpretations, which align with the results derived from the observed data.

To account for the third source of survival bias, we divided the GWAS sample into 10 birth cohorts, segmented by decades. According to the Canadian CF registry 2020 [25], there are significant disparities in overall survival probabilities across birth cohorts, with generally higher survival probability observed in younger cohorts. Within each cohort, we assume that individuals share similar survival rates, thus making stratified analyses less prone to survival bias. We proceeded to conduct stratified analyses at the two genome-wide significant SNPs: rs62369766 (chr5p12) and rs927553 (chr13q12.12). For both SNPs, we noted consistent effect directions across the different birth cohorts. This observation suggests that the effect estimations at these two SNPs are robust.

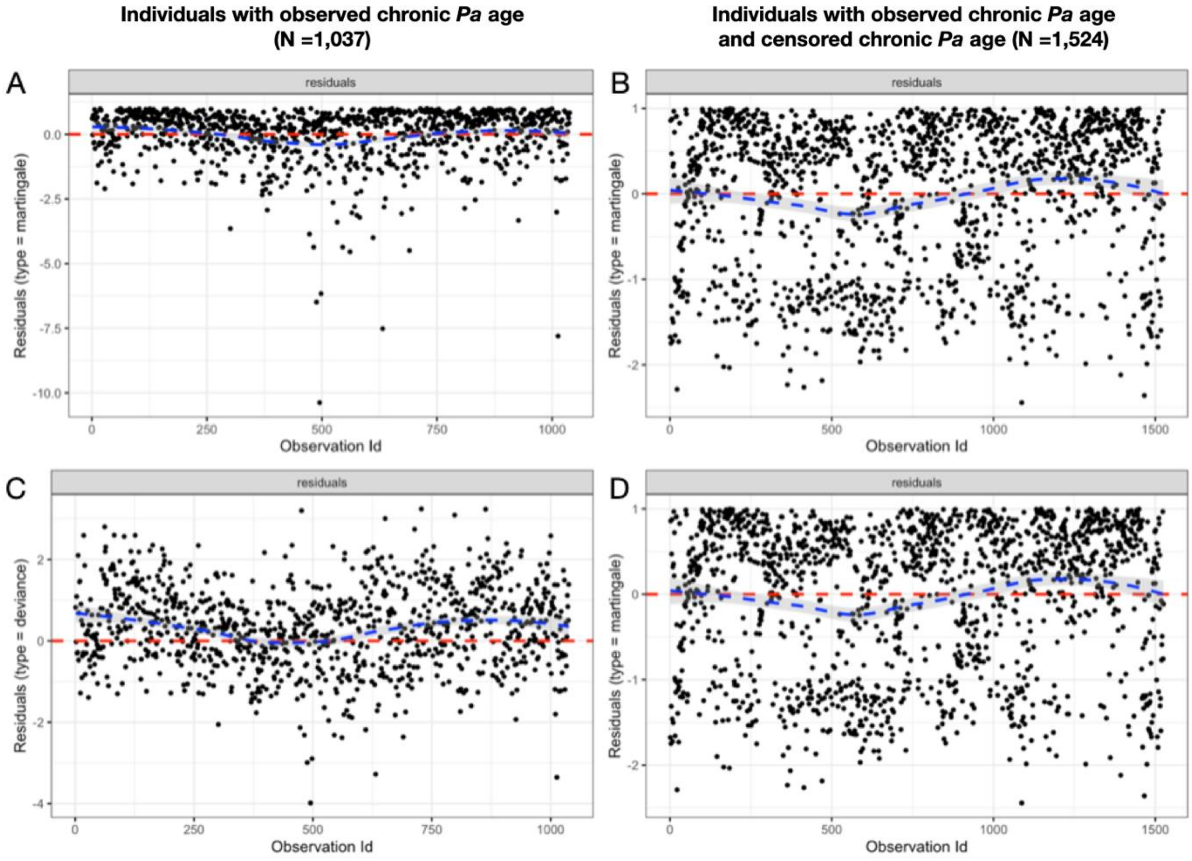


Figure ii: Martingale residuals (A and B) and deviance residuals (C and D) for the Time-to-Event modeling. This figure includes individuals with observed chronic *Pa* age (A and C) and additional individuals with infection age censored at the last longitudinal visit (B and D).

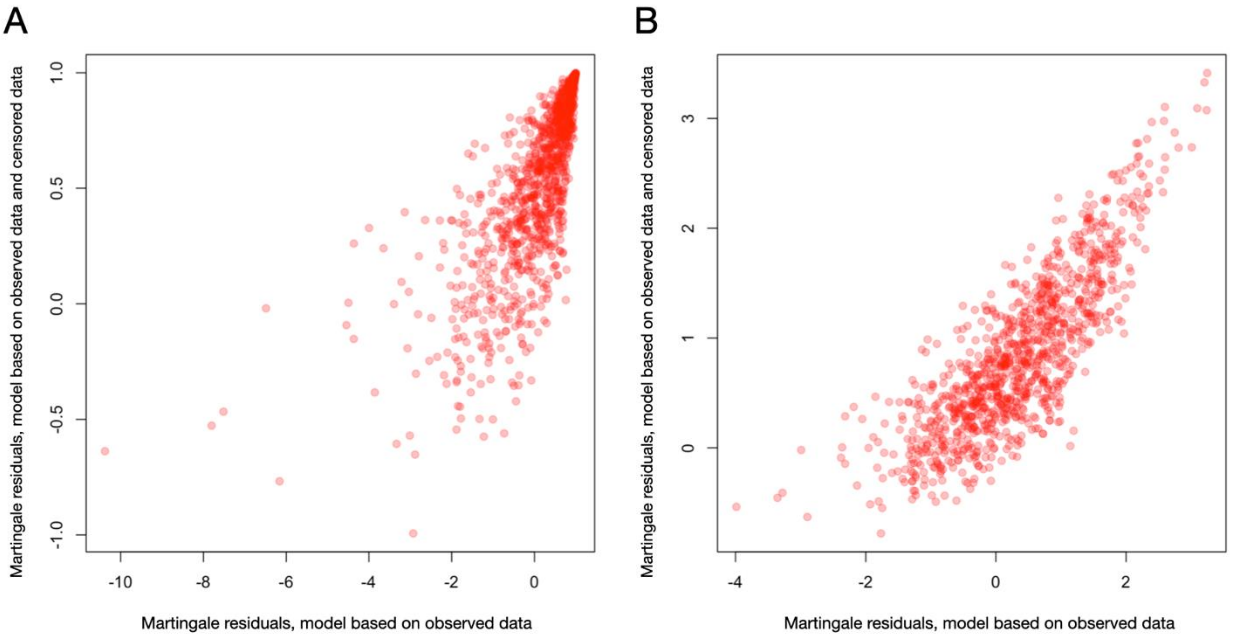


Figure iii: Comparison of Martingale residuals (A) and Deviance residuals (B) for the Time-to-Event model. The x-axis represents the model fitted with individuals who have observed chronic *Pa* age (N=1,037), while the y-axis represents the model fitted with these individuals plus additional individuals whose infection age was censored at their last longitudinal visit (N=1,524). Each dot represents one of the 1,037 individuals with observed chronic *Pa* age.

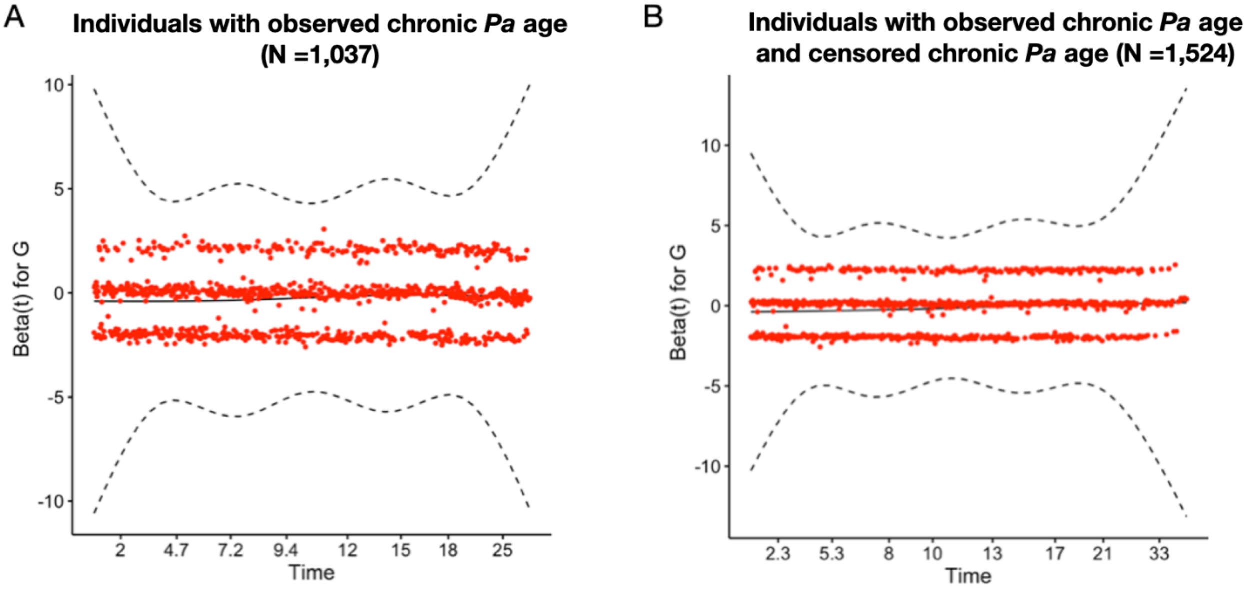


Figure iv: Schoenfeld residuals for the Time-to-Event model. (A) is the model fitted with individuals who have observed chronic *Pa* age (N=1,037), while (B) is the model fitted with these individuals plus additional individuals whose infection age was censored at their last longitudinal visit (N=1,524).

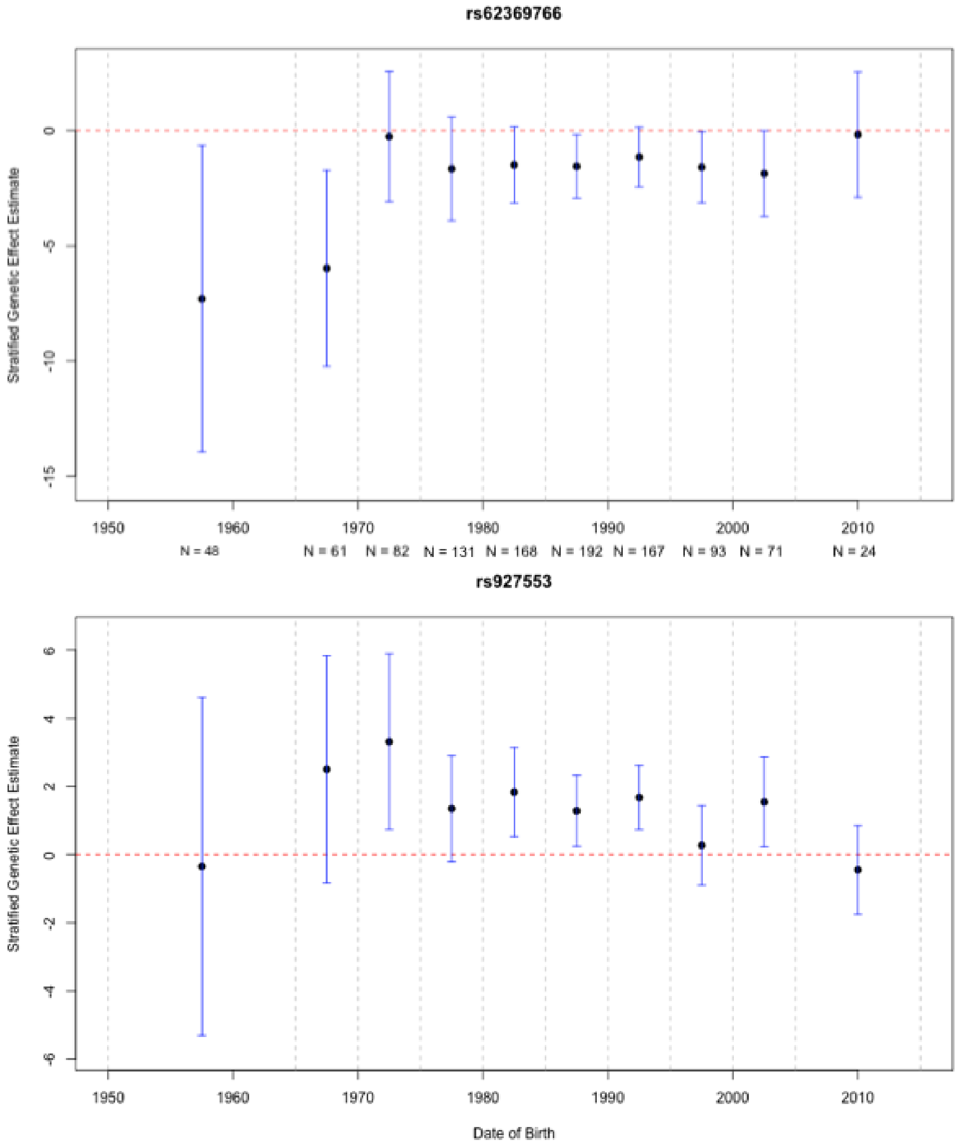


Figure v: Stratified analysis by birth cohort in decade increments for the two genome-wide significant SNPs: rs62369766 (chr5p12) and rs927553 (chr13q12.12). The bars indicate 95% confidence intervals.

Supplementary Appendix 7
Birth-cohort effect adjustment in different analysis

In the GWAS, current age (difference in years between the analysis endpoint and date of birth) was used as a quantitative covariate to define the residual of the chronic *Pa* age, which served as the response variable in association testing (Supplementary Appendix 1, Figure v). No other birth cohort variables were further adjusted for.

In the TTE analysis incorporating individuals with censored infection ages, the age was included as a covariate in the Cox model, together with other covariates including sex and PCs.

In the sensitivity analysis accounting for changing treatment over eras (Supplementary Appendix 5, Table iii), we stratified 1,037 GWAS individuals into the three broader birth cohorts and replicated the same association tests as the one used in GWAS within each birth cohort. (Supplementary Appendix 5)

In the sensitivity analysis checking potential survival bias (Supplementary Appendix 6, Figure v), we divided the GWAS sample into 10 birth cohorts segmented by decades and checked effect directions across these 10 birth cohorts.

Web Resources

LocusZoom, http://locuszoom.org/

GENESIS, https://bioconductor.org/packages/release/bioc/html/GENESIS.html

KING, https://www.kingrelatedness.com/

GWAS atlas, https://atlas.ctglab.nl/

Genotype-Tissue Expression Project (GTEx) Portal, https://www.gtexportal.org/home/

PRSice-2, https://www.prsice.info/

PLINK 1.9, https://www.cog-genomics.org/plink/

LocusFocus, https://locusfocus.research.sickkids.ca/

Reference

1. Corvol, H., S.M. Blackman, P.Y. Boelle, et al., *Genome-wide association meta-analysis identifies five modifier loci of lung disease severity in cystic fibrosis.* Nat Commun, 2015. **6**: p. 8382.

2. Gong, J., F. Wang, B. Xiao, et al., *Genetic association and transcriptome integration identify contributing genes and tissues at cystic fibrosis modifier loci.* PLoS Genet, 2019. **15**(2): p. e1008007.

3. Anderson, C.A., F.H. Pettersson, G.M. Clarke, et al., *Data quality control in genetic case-control association studies.* Nat Protoc, 2010. **5**(9): p. 1564-73.

4. Manichaikul, A., J.C. Mychaleckyj, S.S. Rich, et al., *Robust relationship inference in genome-wide association studies.* Bioinformatics, 2010. **26**(22): p. 2867-73.

5. Team, R.D.C., *R: A Language and Environment for Statistical Computing*. 2010, R Foundation for Statistical Computing.

6. Gogarten, S.M., T. Sofer, H. Chen, et al., *Genetic association testing using the GENESIS R/Bioconductor package.* Bioinformatics, 2019. **35**(24): p. 5346-5348.

7. Genomes Project, C., A. Auton, L.D. Brooks, et al., *A global reference for human genetic variation.* Nature, 2015. **526**(7571): p. 68-74.

8. Panjwani, N., B. Xiao, L. Xu, et al., *Improving imputation in disease-relevant regions: lessons from cystic fibrosis.* NPJ Genom Med, 2018. **3**: p. 8.

9. Browning, B.L. and S.R. Browning, *Genotype Imputation with Millions of Reference Samples.* Am J Hum Genet, 2016. **98**(1): p. 116-26.

10. Lehmann, E.L., *Elements of large sample theory*. Corr. 2nd prtg. ed. Springer texts in statistics. 2001, New York: Springer.

11. McCaw, Z.R., J.M. Lane, R. Saxena, et al., *Operating characteristics of the rank-based inverse normal transformation for quantitative trait analysis in genome-wide association studies.* Biometrics, 2020. **76**(4): p. 1262-1272.

12. Green, D.M., J.M. Collaco, K.E. McDougal, et al., *Heritability of respiratory infection with Pseudomonas aeruginosa in cystic fibrosis.* J Pediatr, 2012. **161**(2): p. 290-5 e1.

13. Yang, J., S.H. Lee, M.E. Goddard, et al., *GCTA: a tool for genome-wide complex trait analysis.* Am J Hum Genet, 2011. **88**(1): p. 76-82.

14. Watanabe, K., S. Stringer, O. Frei, et al., *A global overview of pleiotropy and genetic architecture in complex traits.* Nat Genet, 2019. **51**(9): p. 1339-1348.

15. Shrine, N., A.L. Guyatt, A.M. Erzurumluoglu, et al., *New genetic signals for lung function highlight pathways and chronic obstructive pulmonary disease associations across multiple ancestries.* Nat Genet, 2019. **51**(3): p. 481-493.

16. Consortium, G.T., *The Genotype-Tissue Expression (GTEx) project.* Nat Genet, 2013. **45**(6): p. 580-5.

17. Hemani, G., J. Zheng, B. Elsworth, et al., *The MR-Base platform supports systematic causal inference across the human phenome.* Elife, 2018. **7**.

18. Choi, S.W., T.S. Mak, and P.F. O'Reilly, *Tutorial: a guide to performing polygenic risk score analyses.* Nat Protoc, 2020. **15**(9): p. 2759-2772.

19. Burgess, S. and S.G. Thompson, *Interpreting findings from Mendelian randomization using the MR-Egger method.* Eur J Epidemiol, 2017. **32**(5): p. 377-389.

20. Lawlor, D.A., R.M. Harbord, J.A. Sterne, et al., *Mendelian randomization: using genes as instruments for making causal inferences in epidemiology.* Stat Med, 2008. **27**(8): p. 1133-63.

21. Hartwig, F.P., N.M. Davies, and G. Davey Smith, *Bias in Mendelian randomization due to assortative mating.* Genet Epidemiol, 2018. **42**(7): p. 608-620.

22. Burgess, S., R.A. Scott, N.J. Timpson, et al., *Using published data in Mendelian randomization: a blueprint for efficient identification of causal risk factors.* Eur J Epidemiol, 2015. **30**(7): p. 543-52.

23. Bowden, J., G. Davey Smith, P.C. Haycock, et al., *Consistent Estimation in Mendelian Randomization with Some Invalid Instruments Using a Weighted Median Estimator.* Genet Epidemiol, 2016. **40**(4): p. 304-14.

24. Hartwig, F.P., G. Davey Smith, and J. Bowden, *Robust inference in summary data Mendelian randomization via the zero modal pleiotropy assumption.* Int J Epidemiol, 2017. **46**(6): p. 1985-1998.

25. Canada, C.F., *The Canadian Cystic Fibrosis Registry 2020 Annual Data Report*. 2022.