Genetic modifiers of cystic fibrosis lung disease severity: whole genome analysis of 7,840 patients

Yi-Hui Zhou, PhD,[1, 2,*] Paul J. Gallins, MS,[1] Rhonda G. Pace, BS,[3] Hong Dang, PhD,[3] Melis A. Aksit, PhD,[4] Elizabeth E. Blue, PhD,[5, 6] Kati J. Buckingham, BS,[7] Joseph M. Collaco, MD, MS, PhD,[8] Anna V. Faino, MS,[9] William W. Gordon, MS,[7] Kurt N. Hetrick, MS,[10] Hua Ling, PhD,[10] Weifang Liu, BA,[11] Frankline M. Onchiri, PhD,[9] Kymberleigh Pagel, PhD,[12] Elizabeth W. Pugh, PhD,[10] Karen S. Raraigh, MGC, CGC,[4] Margaret Rosenfeld, MD, MPH,[13, 14] Quan Sun, BS,[11] Jia Wen, PhD,[15] Yun Li, PhD,[11, 15, 16] Harriet Corvol, PhD,[17, 18] Lisa J. Strug, PhD,[19, 20, 21, 22, 23] Michael J. Bamshad, MD,[5, 7, 24] Scott M. Blackman, MD, PhD,[4, 25] Garry R. Cutting, MD,[4] Ronald L. Gibson, MD, PhD,[13, 14] Wanda K. O'Neal, PhD,[3,27] Fred A. Wright, PhD,[1, 2, 26,27] Michael R. Knowles, MD[3,27,*] on behalf of the Cystic Fibrosis Genome Project

[1]Bioinformatics Research Center, North Carolina State University, Raleigh, North Carolina, 27695, USA

[2]Department of Biological Sciences, North Carolina State University, Raleigh, North Carolina, 27695, USA

[3]Marsico Lung Institute/UNC CF Research Center, School of Medicine, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, 27599, USA

[4]McKusick-Nathans Department of Genetic Medicine, Johns Hopkins University School of Medicine, Baltimore, Maryland, 21287, USA

[5]Brotman Baty Institute for Precision Medicine, Seattle, Washington, 98195, USA

[6]Department of Medicine, Division of Medical Genetics, University of Washington, Seattle, Washington, 98195, USA

AJRCCM Articles in Press. Published March 15, 2023 as 10.1164/rccm.202209-1653OC

[7]Department of Pediatrics, Division of Genetic Medicine, University of Washington, Seattle, Washington, 98195, USA

[8]Eudowood Division of Pediatric Respiratory Sciences, Johns Hopkins University School of Medicine, Baltimore, Maryland, 21287, USA

[9]Children's Core for Biostatistics, Epidemiology and Analytics in Research, Seattle Children's Research Institute, Seattle, Washington, 98101, USA

[10]Department of Genetic Medicine, Center for Inherited Disease Research, Johns Hopkins University School of Medicine, Baltimore, Maryland, 21205, USA

[11]Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, 27599, USA

[12]The Institute for Computational Medicine, The Johns Hopkins University, Baltimore, Maryland, 21218, USA

[13]Center for Clinical and Translational Research, Seattle Children's Research Institute, Seattle, Washington, 98101, USA

[14]Department of Pediatrics, University of Washington, Seattle, Washington, 98195, USA

[15]Department of Genetics, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, 27599, USA

[16]Department of Computer Science, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, 27599, USA

[17]Assistance Publique-Hôpitaux de Paris, Hôpital Trousseau, Pediatric Pulmonary Department, Paris, France

AJRCCM Articles in Press. Published March 15, 2023 as 10.1164/rccm.202209-1653OC

[18]Sorbonne Université, Institut National de la Santé et de la Recherche Médicale, Centre de Recherche Saint Antoine, Paris, France

[19]Division of Biostatistics, Dalla Lana School of Public Health, University of Toronto, Toronto, ON, Canada

[20]Program in Genetics and Genome Biology, The Hospital for Sick Children, Toronto, ON, Canada

[21]Department of Statistical Sciences, University of Toronto, Toronto, ON, Canada

[22]The Center for Applied Genomics, The Hospital for Sick Children, Toronto, ON, Canada

[23]Department of Computer Science, University of Toronto, Toronto, ON, Canada.

[24]Department of Genome Sciences, University of Washington, Seattle, Washington, 98195, USA

[25]Division of Pediatric Endocrinology, Johns Hopkins University School of Medicine, Baltimore, Maryland, 21287, USA

[26]Department of Statistics, North Carolina State University, Raleigh, North Carolina, 27695, USA

[27]These authors contributed equally

*Correspondence: Michael R. Knowles, MD, Division of Pulmonary Diseases and Critical Care Medicine, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599; email: knowles@med.unc.edu and Yi-Hui Zhou, PhD, Bioinformatics Research Center and Department of Biological Sciences, North Carolina State University, Raleigh, North Carolina, 27695, USA; email: yihui_zhou@ncsu.edu.

Author Contributions: R.G.P, E.E.B., J.M.C., M.R.K: data acquisition/data analysis/interpretation/final manuscript review; M.A.A., A.V.F., F.O., K.S.R.: data acquisition/data analysis/final manuscript review; G.R.C., R.L.G.: data acquisition/interpretation/final manuscript review; K.J.B., M.R, H.C., L.J.S., M.J.B., S.M.B: data

AJRCCM Articles in Press. Published March 15, 2023 as 10.1164/rccm.202209-1653OC

acquisition/final manuscript review; W.W.G., K.N.H., H.L., W.L., K.P., Q.S., J.W., Y.L.: data analysis/final manuscript review; Y-H.Z., P.J.G., F.A.W., H.D., E.W.P.: data analysis/interpretation/final manuscript review; Y-H.Z., P.J.G., H.D., Y.L., F.A.W.: statistical analysis; Y-H.Z., P.J.G., R.G.P., H.D., F.A.W., W.K.O., M.R.K.: manuscript drafting; Y-H.Z., F.A.W., W.K.O., M.R.K: study design; H.C., L.J.S., M.J.B., S.M.B., G.R.C., R.L.G, M.R.K.: obtained funding

At a Glance Commentary

Current scientific knowledge on the subject: Genetic modifiers affect lung disease severity.

What this study adds to the field: This is the largest pre-modulator GWAS to date for lung function in CF, using recently generated WGS. Genetic variation controlling CF lung disease severity spans the biological spectrum, from innate immunity to lung development, and defines key genes and pathways for future exploration.

Supported by the Cystic Fibrosis Foundation (CUTTIN18XX1, BAMSHA18XX0, KNOWLE18XX0), Canadian Institutes of Health Research (FRN 167282), and Cystic Fibrosis Canada (2626). Was additionally supported by NIH/NIDDK P30 DK065988 and CFF BOUCHE19R0. Support provided by NHLBI, through the BioData Catalyst program (award 1OT3HL142479-01, 1OT3HL142478-01, 1OT3HL142481-01, 1OT3HL142480-01, 1OT3HL147154). This work was also funded by the Government of Canada through Genome Canada (OGI-148) and supported by a grant from the Government of Ontario. Any opinions are

AJRCCM Articles in Press. Published March 15, 2023 as 10.1164/rccm.202209-1653OC

those of the authors and do not reflect the views of NHLBI, individual BioData Catalyst members, or affiliated organizations.

Short running head: Genetic modifiers of CF lung disease severity

Category: 9.16 Cystic Fibrosis: Basic Science

Word count: 3797

This article has an online supplement, which is accessible from this issue's table of contents online at www.atsjournals.org.

AJRCCM Articles in Press. Published March 15, 2023 as 10.1164/rccm.202209-1653OC

ABSTRACT

Rationale: Lung disease is the major cause of morbidity and mortality in persons with cystic fibrosis (pwCF). Variability in CF lung disease has substantial non-CFTR genetic influence. Identification of genetic modifiers has prognostic and therapeutic importance.

Objectives: Identify genetic modifier loci and genes/pathways associated with pulmonary disease severity.

Methods: Whole genome sequencing (WGS) data on 4,248 unique pwCF with pancreatic insufficiency (PI) and lung function measures were combined with imputed genotypes from an additional 3,592 PI patients from the US, Canada, and France. This report describes association of ~15.9 million single nucleotide polymorphisms (SNPs), using the quantitative Kulich Normal Residual Mortality Adjusted (KNoRMA) lung disease phenotype in 7,840 pwCF using pre-modulator lung function data.

Measurements and Main Results: Testing included common and rare SNPs, transcriptome-wide association, gene level, and pathway analyses. Pathway analyses identified novel associations with genes that have key roles in organ development, and we hypothesize these genes may relate to dysanapsis and/or variability in lung repair. Results confirmed and extended previous GWAS findings. These WGS data provide finely mapped genetic information to support mechanistic studies. No novel primary associations with common single variants or with rare variants were found. Multi-locus effects at chr5p13 (*SLC9A3/CEP72*) and chr11p13 (*EHF/APIP*) were identified. Variant effect size estimates at associated loci were consistently ordered across the cohorts, indicating possible age or birth cohort effects.

AJRCCM Articles in Press. Published March 15, 2023 as 10.1164/rccm.202209-1653OC

Conclusions: This pre-modulator genomic, transcriptomic, and pathway association study of 7,840 pwCF will facilitate mechanistic and post-modulator genetic studies and, development of novel therapeutics for CF lung disease.

There are 250 words in this abstract.

Keywords: cystic fibrosis, whole genome sequencing, lung disease severity, GWAS/TWAS, pathway analyses

AJRCCM Articles in Press. Published March 15, 2023 as 10.1164/rccm.202209-1653OC

INTRODUCTION

Lung disease is the major cause of morbidity and mortality in people with cystic fibrosis (pwCF; CF) (1), but the severity can vary widely among individuals. In part, this variation reflects genetic variants in cystic fibrosis transmembrane conductance regulator (*CFTR*) (2) that span a spectrum of severity, from complete loss-of-function (LOF) mutations that are associated with exocrine pancreatic insufficiency (PI), to *CFTR* variants with residual function (2). Additionally, while environmental influences contribute to lung disease variability, non-*CFTR* modifier genes also play a role (heritability 0.54) (3, 4). While recent advances in CFTR modulator therapies have improved outcomes for many pwCF, some do not benefit due to non-responsive genetic variants in *CFTR*. Continued exploration of non-*CFTR* genetic modifiers is expected to provide new therapeutic targets (5).

A previously published genome-wide association study (GWAS) for lung disease severity in pwCF with PI reported modifier variants at five loci (6), with an additional significant GWAS locus identified using improved imputation of SNP genotypes from the primary paper (7). These previous studies utilized whole genome SNP arrays and a validated lung disease phenotype, Kulich Normal Residual Mortality Adjusted (KNoRMA), which is based on multiple measurements of forced expiratory volume in one second (FEV_1), corrected for sex, age and survival, enabling analysis across different ages and cohorts (6).

Whole genome sequencing (WGS) has made it possible to study genotype-phenotype associations at high resolution. The Cystic Fibrosis Genome Project (CFGP) is a multi-site consortium to dissect molecular sources for the variability of phenotypes in pwCF (8). We reasoned that combining data from WGS samples with samples and data from prior GWAS would provide a highly resolved picture of CF lung disease phenotype-genotype associations and

AJRCCM Articles in Press. Published March 15, 2023 as 10.1164/rccm.202209-1653OC

a more detailed biological understanding of CF lung disease. We report extensive analyses using KNoRMA, calculated from lung function data prior to modulator therapy, to identify genetic modifiers of pulmonary disease in 7,840 pwCF, the largest such study to date. These rigorous pre-modulator data will inform both ongoing therapeutic development studies and serve as a basis for future post-modulator genome studies.

Some of the results of these studies have been previously reported in the form of an abstract (9).

METHODS

The online **Supplement** includes numerous details, with brief descriptions provided here.

A total of 5,199 CFGP samples were sequenced (8). Of those, 4,248 were PI patients with sufficient pre-modulator lung function measures for inclusion from three studies/sites: the Gene Modifier Study (GMS)/University of North Carolina (UNC); the Twin & Sibling Study (TSS) and CF-Related Diabetes Studies/Johns Hopkins University (JHU); and the Early Pseudomonas Infection Control Study (EPIC)/University of Washington (UW) (**Table 1**). These WGS data were combined with an independent set of 3,592 patients with genome-wide genotypes imputed from TOPMed data (10) from array-based genotypes (6) (total = 7,840).

A validated quantitative lung function trait was calculated using the KNoRMA phenotype (6), which allows analyses across age, gender, and cohorts. KNoRMA is based on multiple measures of FEV_1 over three years using data from the CF Foundation Patient Registry (CFFPR

AJRCCM Articles in Press. Published March 15, 2023 as 10.1164/rccm.202209-1653OC

2017) (11), and is corrected for age, sex, and survival. A disease progression and mortality-adjusted phenotype such as KNoRMA increases statistical power while reducing the need for stratification or additional covariates. For this study, in order to avoid the confounding effects of recently approved modulators, KNoRMA was calculated from FEV_1 prior to modulator therapy (see **Supplement** for details).

CFGP samples were sequenced to ~30X coverage with careful quality/identity checks (see details in **Supplement**). The GWAS array-based data and cohorts were described previously (6). Genetic imputation for the non-CFGP samples was performed as described (10).

Analyses used a quantitative trait of lung disease severity (KNoRMA) (6). The primary analyses included non-rare single-variant SNP testing (minor allele count ≥ 20). Association was tested using KNoRMA as a response in an additive effect mixed model, using ancestry, sex, and terms for site×platform combinations as covariates. A genetic relatedness matrix was used to account for the small proportion of families and cryptic relatedness. Results were combined across site×platform as a fixed-effect meta-analysis. *P* value thresholds were applied at the genome-wide significance level ($P < 5 \times 10^{-8}$) (12), and we considered SNPs with $P < 5 \times 10^{-7}$ to be suggestive.

For the significant GWAS loci, we ran CAVIAR (Causal Variants Identification in Associated Regions), to assess evidence of SNP causality (13). The Ensembl Variant Effect Predictor (VEP) was used to determine putative effects of variants on genes, transcripts, protein sequences, and regulatory regions (14).

Transcriptome-wide association (TWAS) evidence was determined from 50 tissues, using a summary association z-statistic (15). This approach uses SNP-level gene expression weights

AJRCCM Articles in Press. Published March 15, 2023 as 10.1164/rccm.202209-1653OC

from 48 tissues from the Genotype–Tissue Expression (GTEx) project v8 (16), peripheral blood from Netherlands Twin Registry (NTR) (17), and whole blood from the Young Finns Study (YFS) (18). For these and all gene-based approaches, including gene association summaries and rare-variant methods, we used a false discovery q value < 0.1 to declare significance.

Although the KNoRMA phenotype is corrected for age-dependent effects on survival, we additionally devised a reverse regression approach to investigate potential age interactions for genotype associations (**Supplemental** Methods). In addition, a method was devised to assess concordant effect size ordering across site cohorts for different loci, using a summary of pairwise correlations of estimated effect sizes across cohorts, with statistical significance assessed by permutation. For our meta-analysis statistic, we show that this permutation approach remains valid under selection for genome-wide significance (**Supplemental Methods**). Gene-level summary analyses were performed using VEGAS2 (19) for intragenic SNPs and SNPs within a flanking region of 20kb around each gene. Gene-based pathway analyses were performed using Gene Set Enrichment Analysis (GSEA) method (20), available in the clusterProfiler R package (21). Rare variant methods (minor allele count < 20) were performed at the gene level using the GENESIS R package for the burden test, SMMAT, and SKAT-O.

RESULTS

Table 1 describes key features of the five cohorts, including the country of enrollment. The majority (n = 4,248) of these PI pwCF had WGS, and the remainder (n = 3,592) had genotypes imputed from WGS (10). The means and standard deviations for lung disease severity (KNoRMA) were comparable across cohorts, despite considerable differences in mean (and

AJRCCM Articles in Press. Published March 15, 2023 as 10.1164/rccm.202209-1653OC

median) age. The UW cohort includes the youngest patients (mean 13.1 years), while UNC was the oldest cohort (mean 26.8 years). The vast majority of these pwCF are of European ancestry (~95%), and most (~64%) are c.1521_1523del (p.Phe508del; legacy: F508del) homozygotes. Effective ancestry control could be achieved by four genotype principal components (22) and we used six principal components to be conservative. The violin plot illustrates the distribution of KNoRMA by cohort, and the UNC plot shows two distinct modes, reflecting an extremes-of-phenotype design (**Figure 1**) (6).

The GWAS analysis for KNoRMA using non-rare variants (minor allele count > 20, hereafter termed "common"), identified six genome-wide significant ($P < 5 \times 10^{-8}$) (12) loci (**Figure 2**, **Table 2**, **Figures E1** and **E2)**. The present analyses increased the significance of 4 of 5 loci reported in our previous GWAS (**Table 2**) (6). The sixth locus at 16p12.2 near *CHP2* and *PRKCB* ($P = 2.5 \times 10^{-8}$) (**Figure E2**, **Table 2**) was not reported in our previous GWAS (6), but was identified in a separate analysis using updated and improved imputation of SNP genotypes (7). Each of these six loci contains genes of high biological relevance to the pathophysiology of CF lung disease (4, 7). Four suggestive loci ($P < 5 \times 10^{-7}$; all with low MAF, range 0.005 to 0.009) were also identified, including chr1p36 (*CEP85*), chr6q15 (*UBE2J1*), chr8q11.2 (*SNTG1*), and chr17q22 (*PPM1E*) (**Figure E3**, **Table 2,** and **Supplement**). Finally, we identified many associations (P value $< 10^{-5}$) with KNoRMA in all (7,840) pwCF and 4,985 F508del homozygotes (**Table E1**).

Conditioning on the top-ranked SNP in six regions with genome-wide significance eliminated significant secondary signals in four regions, but two loci (chr5p15.33; *SLC9A3/CEP72* and chr11p13; *APIP*/*EHF*) displayed regional significance for secondary SNPs (**Figures E2**). By fitting all regional two SNP models for chr5p15 and chr11p13 (see **Methods**),

AJRCCM Articles in Press. Published March 15, 2023 as 10.1164/rccm.202209-1653OC

we determined the best-fitting SNP pair for each region. For chr5p15, conditioning on the primary SNP (rs56108664) revealed a significant secondary SNP (rs111275646) and other SNPs in linkage disequilibrium (LD) (**Figure E4**). Conditioning on the secondary SNP at chr5p15 recapitulated the original signal (conditional *P* value for the original SNP rs56108664 was $P \sim 3 \times 10^{-8}$ after conditioning on rs111275646). For the chr11p13 locus, the use of the best two-SNP model (primary, rs483769; secondary, rs1509661) provided informative results (**Figure E5**). Namely, the *P* values for SNPs in the primary LD "block" after conditioning on the secondary SNP (rs1509661) became ~10,000-fold smaller ($P \sim 7 \times 10^{-14}$) than in the original single variant analysis ($P \sim 2.6 \times 10^{-9}$) (**Table 2**). Further investigation of the chr11p13 locus revealed that the minor alleles of the primary and secondary SNP are positively associated ($r^2 = 0.28$) but have associations in opposite directions with KNoRMA. In this scenario, most subjects in this study have at least one risk allele at the primary locus and at least one protective allele at the secondary locus, with combinations of risk alleles from either locus contributing to overall phenotype consequences (**Figure E6**). For the chr5p15 and chr11p13 regions, haplotype analyses that account for linkage phase (**Supplementary Results**) were not more significant than the primary genotype-based analyses.

Causality at each significant locus has not been established due to LD structure, and analysis by Causal Variants Identification in Associated Regions (CAVIAR) (13) and annotation of the top SNPs in the six regions with Variant Effect Predictor (VEP) (14) did not point to any obvious causal links (**Table E2**). The gene-level rare variant analyses did not identify any significant gene at $q < 0.1$, perhaps reflecting reduced power for rare variant detection compared to common variant analyses.

AJRCCM Articles in Press. Published March 15, 2023 as 10.1164/rccm.202209-1653OC

Our reverse regression model included terms for age at phenotyping and age×KNoRMA interaction, and largely recapitulated our main findings, with five of the six reported loci achieving significance (rs194788 near *CHP2* achieving only $P = 3.23\times10^{-7}$), and no new significant findings. The age and age×KNoRMA interaction terms were not significant for these regions. Nonetheless, substantial cohort variation was apparent. The effect sizes (magnitude of beta coefficients) for the peak SNP at the six significant loci were evaluated using forest plots for each cohort, ordered by mean age (**Figure 3**). There is a similar distribution of the effect (size) across cohorts, with UNC cohort (oldest) showing the largest effect size and UW (youngest) showing the smallest. This concordance of effect sizes manifests as positive correlation in all pairwise comparisons (mean correlation 0.70), as depicted in plots of effect sizes (**Figure E7**). A test of concordance of effect sizes demonstrated concordance among cohorts ($P = 3.4 \times 10^{-4}$), with the UW cohort consistently exhibiting the smallest effect size.

To further investigate genotype associations for all pwCF (n = 7,840), we imputed expression values to investigate TWAS association with lung phenotype, using a modified approach (15) compared to a previous study of TWAS in CF (7). Twenty-nine annotated genes displayed a false-discovery $q < 0.1$ (**Figure 4**; **Tables E3** and **E4**). Most genes with significant TWAS signals occurred in the six significant GWAS loci (**Figure 2**), congruent with a previous report (7). In this analysis, *MUC4* was suggestive ($q = 0.14$).

Previous studies suggested that there may be GWAS loci associated more strongly with pwCF homozygous for the *CFTR* variant F508del (6). Analyses in this study identified three new suggestive loci (**Figures E8 – E10**).

Significant results ($q < 0.10$) from the VEGAS2 gene-level association analyses are provided (**Table E5**). As gene-level analyses can capture effects of long-range linkage

AJRCCM Articles in Press. Published March 15, 2023 as 10.1164/rccm.202209-1653OC

disequilibrium, we grouped significant regions into those separated by more than 5 Mb. Five of the six regions with individually-significant SNPs (see above) were also significant in gene-level analysis (excepting the chr11p13 region). Among the remaining significant genes identified by VEGAS2, several achieved Bonferroni significance at a more stringent $\alpha = 0.05$ ($P < 2.5 \times 10^{-6}$): *ADAMTS8*, *LINC01844*, and *PTTG1IP*.

We performed GSEA on genes ranked using VEGAS2 *P* values (**Table E5**) (19) to explore pathways linked to lung disease severity. Pathways identified were largely related to pathogenic mechanisms linked to pulmonary host defense and genes at GWAS significant loci (**Table E6**) (6, 7, 23, 24), involving: inflammation; viral and bacterial infection and host responses; immunity and HLA-II pathways; endomembrane function; and microtubular/cytoskeletal function. In addition, multiple pathways related to organ development and morphogenesis were identified (**Table E6**). The most significant development/morphogenesis pathway [GO.BP0048754, branching morphogenesis of an epithelial tube; GSEA plot (**Figure 5**)] includes 32 genes in the leading edge (in bold) that relate to three signaling pathways (Sonic Hedgehog, Shh; transforming growth factor beta, TGFb; wingless related-integration site (Wnt)/β-catenin) that are necessary for lung development and branching morphogenesis (25-27). Thus, genetic variation that affects lung development in utero and early childhood has implications for severity of CF lung disease.

DISCUSSION

Variability of lung disease severity in CF reflects substantial non-*CFTR* genetic variation (3). Identifying the molecular basis of CF lung disease severity will provide pathobiological

AJRCCM Articles in Press. Published March 15, 2023 as 10.1164/rccm.202209-1653OC

insights and identify new therapeutic targets. Studies using genotype array-based platforms and a standardized lung disease phenotype (KNoRMA) in different cohorts, study designs, and ages/birth cohorts, have identified non-*CFTR* genetic variation of high mechanistic interest (6). By combining WGS with imputation from array-based genotypes across multiple cohorts, we provide the largest analyses to date associating genetic variation to CF lung disease severity in the pre-modulator era to date (7,840 pwCF; an estimated 19% of the number of CF patients currently in North America and France) (2, 4).

One novel insight emerged from pathway analyses (GSEA) of genes ranked by VEGAS2, where multiple significant pathways related to organ development were identified. While not annotated specifically to the lung, the genes within these pathways, especially those related to three key signaling pathways (Shh, TGFb, and Wnt), are known to be critical for lung development and branching morphogenesis (25-27). There are at least 40 genes that relate to these three key signaling pathways in the top annotated pathway (**Figure 5**). The next challenge will be to decipher the mechanism by which these genes could influence CF lung disease severity. The issue is complex because not only do these genes play a role in lung development/morphogenesis, but it is also now appreciated that reactivation of developmental genes/pathways is a necessary component of lung repair after injury/inflammation (28). Several potential complementary mechanisms could be operative. First, variable early-life growth of the bronchial tree airway diameter relative to lung volume (dysanapsis) was proposed nearly 50 years ago (29). There is now anatomical evidence from computed tomography to confirm dysanapsis in conducting airways, and presence of smaller diameter bronchi is known to associate with COPD and childhood asthma (30-33). Dysanapsis has not been previously recognized as a potential pathogenic driver of CF airways disease but, given the periods of

AJRCCM Articles in Press. Published March 15, 2023 as 10.1164/rccm.202209-1653OC

bronchial injury common in CF, dysanapsis could have profound effects on long-term outcomes. Second, CFTR itself is known to interact with lung development in several ways: tracheal and proximal bronchial diameter is altered in CF compared to normal pigs during embryonic development (34); CFTR plays a key role in fluid mediated distension of airways during development (35); and lack of CFTR with consequent infections and inflammation are associated with tracheomalacia, which is linked to poorer outcomes (36, 37). Finally, because reactivation of developmental pathways is important for repair after airway injury (28), genetic variation in these pathways is expected to alter outcome after CF-related inflammatory damage. Other significant pathways involve microtubular/cytoskeletal function (see **Figure E11**; GO.BP0051494 "negative regulation of cytoskeleton organization"), which is of particular interest due to a recent potential therapeutic advance by restoration of microtubular dysfunction in CF cells (38).

The five genome-wide significant loci previously reported (6) are highly significant in the present analyses and contain genes of relevance (2, 4). In addition, another locus (chr16p12.2) is genome-wide significant, and four new loci are suggestive ($P < 5x10^{-7}$; all with low MAF, range 0.005 to 0.009). The newly significant locus on chr16p12.2 is intergenic between *CHP2* and *PRKCB* (39, 40). CHP2 regulates airway pH through the apical membrane Na^+/H^+ exchanger (40). A SNP at this locus (rs11646605) is associated with mycobacterium avium complex lung disease in non-CF patients and is an eQTL for *CHP2* in the lung (41). TWAS analyses point to *CHP2* expression as a key candidate in this region (**Figure 4**). PRKCB is a protein kinase that plays a role in multiple cellular functions, including apoptosis and autophagy (39). Finally, a nearby gene (*ERN2*) regulates airway mucin genes (*MUC5B* and *MUC5AC*) (42).

AJRCCM Articles in Press. Published March 15, 2023 as 10.1164/rccm.202209-1653OC

At the chr3q29 locus, *MUC4* and *MUC20* are highly relevant candidate genes, as they play important roles in lung host defense and mucociliary clearance, which are abnormal in CF (2, 4). These WGS data now support *MUC4* as the mechanistic link, with all significant SNPs intragenic to *MUC4*; plus, *MUC4* is supported by a separate study integrating eQTLs and CF GWAS summary statistics using colocalization analysis (43).

Significant SNPs at the chr5p15.3 locus span ~ 300 kb and cover four pertinent genes expressed in respiratory epithelia. Airway surface liquid pH is abnormal in CF and regulated in part by *SLC9A3*, which codes for a Na^+/H+ exchanger (44). Moreover, variable number of tandem repeats (VNTRs) in this region are associated with expression of *SLC9A3* in CF respiratory epithelia (45). The other three genes at this locus (*EXOC3*, *CEP72* and *TPPP*) are involved in cellular microtubular function, which is abnormal in CF (46, 47). These three microtubule-related genes are consistently seen in TWAS-type studies (**Figure 4** and (7)). Resveratrol is an anti-inflammatory polyphenol that is known to activate several pathways relevant to microtubule stability, and it has been recently shown to restore microtubule function and intracellular transport in CF cells (38). Finally, an intergenic SNP (rs11738281) in *CEP72* in our study is associated with airflow obstruction (reduced FEV_1/FVC ratio) in the UK Biobank GWAS (48) and is in LD ($r^2 = 0.61$) with the most significant regional SNP at this locus (chr5p13).

We observed strong gene expression signatures at the chr6p21.3 (HLA Class II) locus, which is associated with many inflammatory and respiratory conditions (49). In addition to TWAS (**Figure 4)**, differential gene expression and biological pathway studies have identified several HLA-II genes associated with CF lung disease (23, 24), as did our pathway analyses (see

AJRCCM Articles in Press. Published March 15, 2023 as 10.1164/rccm.202209-1653OC

online **Supplement**). Functional interpretation of these data is confounded by many polymorphisms and allotypes of genes in this region (50).

Since the chr11p13 locus was first associated with CF lung disease, it has been extensively studied (51, 52). The most significant SNPs are intergenic between *EHF*, an epithelial transcription factor, and *APIP*, an enzyme involved in inflammation through roles in apoptosis and the methionine salvage pathway (4, 24). Conceptually, either of these genes could impact on CF lung disease severity (2, 4). Regulatory regions are in the significant LD block that interacts with *EHF* and nearby *ELF5* (52, 53), but extensive studies have not identified any eQTLs that might drive the phenotype (7, 51, 52). Further, our TWAS analysis (**Figure 4**) produced no signatures that suggest mechanism. Interpretation of this region is further complicated by finding a second group of significantly associated SNPs over *APIP* after conditioning on the top-ranked SNP. The presence of two significant groups of SNPs at this locus implies the risk for each pwCF can be viewed in terms of four (rather than two) alleles, and minor alleles of the primary and secondary SNPs have opposite associations with KNoRMA, and the effect sizes (betas) for the primary and secondary SNPs are different (0.9 and 0.2, respectively). Taken together, these features create a potential complex molecular interplay among four alleles, whereby genotype association of the primary SNP with KNoRMA are affected by the genotypes of the secondary SNP (see **Figure E6**).

The chrXq22-q23 locus contains two genes (*AGTR2*; *SLC6A14*) that are expressed in respiratory epithelia with functions relevant to pathophysiology of CF lung disease. AGTR2 functions in the renin-angiotensin signaling (RAS2) pathway, which is involved in several aspects of lung biology, including inflammation (4). The RAS pathway is altered in CF, and studies in genetically modified mice have therapeutic implications, as deletion and

AJRCCM Articles in Press. Published March 15, 2023 as 10.1164/rccm.202209-1653OC

pharmacologic inhibition of AGTR2 improves several features of lung function in CF mice (54). AGTR2 is also prominent in pathway analyses (**Figure 5** and **Supplement**). *SLC6A14* encodes an amino acid transporter with pleiotropic effects in CF, as it has been linked to lung disease and neonatal intestinal obstruction, but the pathophysiologic mechanisms have not been defined (55, 56).

We noted significant concordance of effect sizes across significant loci among cohorts, with the youngest (UW) cohort showing the smallest effect size, despite medians and distribution of KNoRMA being similar across cohorts. This may reflect smaller effects of variants on lung function (FEV_1) over a shorter time period in younger CF patients. In addition, age at phenotyping is confounded by year of birth cohort, as improvements in treatment (prior to modulators) may have blunted decline in lung function in these younger pwCF. Therefore, it is challenging to define the specific mechanism(s) for smaller effect size in the youngest cohort.

There are several limitations of this study. First, although this is a large sample size for a study of a rare Mendelian disorder, it is likely underpowered to detect rare lung-disease-associated variants. Second, a replication study was not performed, as there is no adequate CF population readily available. Third, we were unable to establish causality at any locus and identification of causal SNPs is complicated by multiple potential modifiers at each locus. Fourth, some potential variants were not fully queried, such as VNTRs and structural variants. Finally, the population studied largely reflects European ancestry and thus important modifier loci present in other populations may have been missed.

In summary, WGS of pwCF enabled accurate genome-wide imputation, which allowed a pre-modulator association study of genetic variants with lung disease severity in 7,840 pwCF. This approach validated previously identified loci, provided better molecular understanding of

AJRCCM Articles in Press. Published March 15, 2023 as 10.1164/rccm.202209-1653OC

significant loci and enabled discovery of new biologically relevant candidate genes and biological pathways, particularly related to lung development. Taken together, these genomic, transcriptional, and pathways data will inform future mechanistic and post-modulator genetic studies and enable development of novel therapeutics for CF lung disease.

AJRCCM Articles in Press. Published March 15, 2023 as 10.1164/rccm.202209-1653OC

Acknowledgements: The authors would like to thank the Cystic Fibrosis Foundation and Cystic Fibrosis Canada for the use of CF Foundation Patient Registry data and Canadian CF Registry data to conduct this study. Additionally, we would like to thank the patients, care providers, and clinic coordinators at CF Centers throughout the United States, Canada, and France for their contributions to both the CF Foundation Patient Registry, the Canadian CF Registry, and the French CF Registry.

Declaration of interests

WKO and JMC serve as grant reviewers for the Cystic Fibrosis Foundation and receive honoraria.

AJRCCM Articles in Press. Published March 15, 2023 as 10.1164/rccm.202209-1653OC

References

1. Cutting GR. Cystic fibrosis genetics: from molecular understanding to clinical application. *Nat Rev Genet* 2015;16:45-56.
2. Paranjapye A, Ruffin M, Harris A, Corvol H. Genetic variation in CFTR and modifier loci may modulate cystic fibrosis disease severity. *J Cyst Fibros* 2020;19 Suppl 1:S10-S14.
3. Vanscoy LL, Blackman SM, Collaco JM, Bowers A, Lai T, Naughton K*, et al.* Heritability of lung disease severity in cystic fibrosis. *Am J Respir Crit Care Med* 2007;175:1036-1043.
4. O'Neal WK, Knowles MR. Cystic fibrosis disease modifiers: complex genetics defines the phenotypic diversity in a monogenic disease. *Annu Rev Genomics Hum Genet* 2018;19:201-222.
5. Egan ME. Cystic fibrosis transmembrane conductance receptor modulator therapy in cystic fibrosis, an update. *Curr Opin Pediatr* 2020;32:384-388.
6. Corvol H, Blackman SM, Boelle PY, Gallins PJ, Pace RG, Stonebraker JR*, et al.* Genome-wide association meta-analysis identifies five modifier loci of lung disease severity in cystic fibrosis. *Nat Commun* 2015;6:8382.
7. Dang H, Polineni D, Pace RG, Stonebraker JR, Corvol H, Cutting GR*, et al.* Mining GWAS and eQTL data for CF lung disease modifiers by gene expression imputation. *PLoS One* 2020;15:e0239189.
8. Raraigh KS, Aksit MA, Hetrick K, Pace RG, Ling H, O'Neal W*, et al.* Complete CFTR gene sequencing in 5,058 individuals with cystic fibrosis informs variant-specific treatment. *J Cyst Fibros* 2022;21:463-470.
9. Zhou Y, Gallins P, Pace R, Dang H, O'Neal W, Li Y*, et al.* Genetic variants that modify severity of CF lung disease: update from the CF genome project. *J Cyst Fibros* 2021;20:S306.

AJRCCM Articles in Press. Published March 15, 2023 as 10.1164/rccm.202209-1653OC

10. Sun Q, Liu W, Rosen JD, Huang L, Pace RG, Dang H*, et al.* Leveraging TOPMed imputation server and constructing a cohort-specific imputation reference panel to enhance genotype imputation among cystic fibrosis patients. *HGG Adv* 2022;3:100090.

11. Knapp EA, Fink AK, Goss CH, Sewall A, Ostrenga J, Dowd C*, et al.* The Cystic Fibrosis Foundation Patient Registry. Design and methods of a national observational disease registry. *Ann Am Thorac Soc* 2016;13:1173-1179.

12. Dudbridge F, Gusnanto A. Estimation of significance thresholds for genomewide association scans. *Genet Epidemiol* 2008;32:227-234.

13. Hormozdiari F, Kostem E, Kang EY, Pasaniuc B, Eskin E. Identifying causal variants at loci with multiple signals of association. *Genetics* 2014;198:497-508.

14. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GR, Thormann A*, et al.* The Ensembl Variant Effect Predictor. *Genome Biol* 2016;17:122.

15. Gusev A, Ko A, Shi H, Bhatia G, Chung W, Penninx BW*, et al.* Integrative approaches for large-scale transcriptome-wide association studies. *Nat Genet* 2016;48:245-252.

16. GTEx Consortium. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* 2020;369:1318-1330.

17. Ligthart L, van Beijsterveldt CEM, Kevenaar ST, de Zeeuw E, van Bergen E, Bruins S*, et al.* The Netherlands Twin Register: longitudinal research based on twin and twin-family designs. *Twin Res Hum Genet* 2019;22:623-636.

18. Mishra BH, Mishra PP, Raitoharju E, Marttila S, Mononen N, Sievänen H*, et al.* Modular genome-wide gene expression architecture shared by early traits of osteoporosis and atherosclerosis in the Young Finns Study. *Sci Rep* 2021;11:7111.

AJRCCM Articles in Press. Published March 15, 2023 as 10.1164/rccm.202209-1653OC

19. Mishra A, Macgregor S. VEGAS2: Software for More Flexible Gene-Based Testing. *Twin Res Hum Genet* 2015;18:86-91.

20. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA*, et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A* 2005;102:15545-15550.

21. Wu T, Hu E, Xu S, Chen M, Guo P, Dai Z*, et al.* clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. *Innovation (Camb)* 2021;2:100141.

22. Kingston H, Stilp AM, Gordon W, Broome J, Gogarten SM, Ling H*, et al.* Accounting for population structure in genetic studies of cystic fibrosis. *HGG Adv* 2022;3:100117.

23. O'Neal WK, Gallins P, Pace RG, Dang H, Wolf WE, Jones LC*, et al.* Gene expression in transformed lymphocytes reveals variation in endomembrane and HLA pathways modifying cystic fibrosis pulmonary phenotypes. *Am J Hum Genet* 2015;96:318-328.

24. Polineni D, Dang H, Gallins PJ, Jones LC, Pace RG, Stonebraker JR*, et al.* Airway mucosal host defense is key to genomic regulation of cystic fibrosis lung disease severity. *Am J Respir Crit Care Med* 2018;197:79-93.

25. Belgacemi R, Danopoulos S, Deutsch G, Glass I, Dormoy V, Bellusci S*, et al.* Hedgehog signaling pathway orchestrates human lung branching morphogenesis. *Int J Mol Sci* 2022;23:5265.

26. Saito A, Horie M, Nagase T. TGF-β signaling in lung health and disease. *Int J Mol Sci* 2018;19:2460.

27. Aros CJ, Pantoja CJ, Gomperts BN. Wnt signaling in lung development, regeneration, and disease progression. *Commun Biol* 2021;4:601.

AJRCCM Articles in Press. Published March 15, 2023 as 10.1164/rccm.202209-1653OC

28. Whitsett JA, Kalin TV, Xu Y, Kalinichenko VV. Building and Regenerating the Lung Cell by Cell. *Physiol Rev* 2019;99:513-554.

29. Green M, Mead J, Turner JM. Variability of maximum expiratory flow-volume curves. *J Appl Physiol* 1974;37:67-74.

30. Smith BM, Kirby M, Hoffman EA, Kronmal RA, Aaron SD, Allen NB*, et al.* Association of dysanapsis with chronic obstructive pulmonary disease among older adults. *JAMA* 2020;323:2268-2280.

31. Vameghestahbanati M, Kirby M, Tanabe N, Vasilescu DM, Janssens W, Everaerts S*, et al.* Central airway tree dysanapsis extends to the peripheral airways. *Am J Respir Crit Care Med* 2021;203:378-381.

32. Smith BM, Traboulsi H, Austin JHM, Manichaikul A, Hoffman EA, Bleecker ER*, et al.* Human airway branch variation and chronic obstructive pulmonary disease. *Proc Natl Acad Sci U S A* 2018;115:E974-E981.

33. Casas M, den Dekker HT, Kruithof CJ, Reiss IK, Vrijheid M, Sunyer J*, et al.* The effect of early growth patterns and lung function on the development of childhood asthma: a population based study. *Thorax* 2018;73:1137-1145.

34. Adam RJ, Abou Alaiwa MH, Bouzek DC, Cook DP, Gansemer ND, Taft PJ*, et al.* Postnatal airway growth in cystic fibrosis piglets. *J Appl Physiol (1985)* 2017;123:526-533.

35. Brennan SC, Wilkinson WJ, Tseng HE, Finney B, Monk B, Dibble H*, et al.* The extracellular calcium-sensing receptor regulates human fetal lung development via CFTR. *Sci Rep* 2016;6:21975.

AJRCCM Articles in Press. Published March 15, 2023 as 10.1164/rccm.202209-1653OC

36. Fischer AJ, Singh SB, Adam RJ, Stoltz DA, Baranano CF, Kao S*, et al.* Tracheomalacia is associated with lower FEV1 and Pseudomonas acquisition in children with CF. *Pediatr Pulmonol* 2014;49:960-970.

37. Wallis C, Alexopoulou E, Antón-Pacheco JL, Bhatt JM, Bush A, Chang AB*, et al.* ERS statement on tracheomalacia and bronchomalacia in children. *Eur Respir J* 2019;54.

38. Lu B, Corey DA, Kelley TJ. Resveratrol restores intracellular transport in cystic fibrosis epithelial cells. *Am J Physiol Lung Cell Mol Physiol* 2020;318:L1145-L1157.

39. Newton AC. Protein kinase C: perfectly balanced. *Crit Rev Biochem Mol Biol* 2018;53:208-230.

40. Di Sole F, Vadnagara K, Moe OW, Babich V. Calcineurin homologous protein: a multifunctional Ca2+-binding protein family. *Am J Physiol Renal Physiol* 2012;303:F165-179.

41. Namkoong H, Omae Y, Asakura T, Ishii M, Suzuki S, Morimoto K*, et al.* Genome-wide association study in patients with pulmonary Mycobacterium avium complex disease. *Eur Respir J* 2021;58:1902269.

42. Chen G, Ribeiro CMP, Sun L, Okuda K, Kato T, Gilmore RC*, et al.* XBP1S regulates MUC5B in a promoter variant-dependent pathway in IPF airway epithelia. *Am J Respir Crit Care Med* 2019;200:220-234.

43. Wang F, Panjwani N, Wang C, Sun L, Strug LJ. A flexible summary statistics-based colocalization method with application to the mucin cystic fibrosis lung disease modifier locus. *Am J Hum Genet* 2022;109:253-269.

AJRCCM Articles in Press. Published March 15, 2023 as 10.1164/rccm.202209-1653OC

44. Pereira SV, Ribeiro JD, Bertuzzo CS, Marson FAL. Association of clinical severity of cystic fibrosis with variants in the SLC gene family (SLC6A14, SLC26A9, SLC11A1 and SLC9A3). *Gene* 2017;629:117-126.

45. Roshandel D, Mastromatteo S, Wang C, Gong J, Thiruvahindrapuram B, Sung WWL*, et al.* A cystic fibrosis lung disease modifier locus harbors tandem repeats associated with gene expression. *medRxiv* 2022:2022.2003.2028.22272580.

46. Rymut SM, Kampman CM, Corey DA, Endres T, Cotton CU, Kelley TJ. Ibuprofen regulation of microtubule dynamics in cystic fibrosis epithelial cells. *Am J Physiol Lung Cell Mol Physiol* 2016;311:L317-327.

47. Rymut SM, Lu B, Perez A, Corey DA, Lamb K, Cotton CU*, et al.* Acetyl-CoA carboxylase inhibition regulates microtubule dynamics and intracellular transport in cystic fibrosis epithelial cells. *Am J Physiol Lung Cell Mol Physiol* 2019;316:L1081-L1093.

48. Kichaev G, Bhatia G, Loh PR, Gazal S, Burch K, Freund MK*, et al.* Leveraging polygenic functional enrichment to improve GWAS power. *Am J Hum Genet* 2019;104:65-75.

49. D'Antonio M, Reyna J, Jakubosky D, Donovan MK, Bonder MJ, Matsui H*, et al.* Systematic genetic analysis of the MHC region reveals mechanistic underpinnings of HLA type associations with disease. *Elife* 2019;8:e48476.

50. Osoegawa K, Mallempati KC, Gangavarapu S, Oki A, Gendzekhadze K, Marino SR*, et al.* HLA alleles and haplotypes observed in 263 US families. *Hum Immunol* 2019;80:644-660.

51. Dang H, Gallins PJ, Pace RG, Guo XL, Stonebraker JR, Corvol H*, et al.* Novel variation at chr11p13 associated with cystic fibrosis lung disease severity. *Hum Genome Var* 2016;3:16020.

AJRCCM Articles in Press. Published March 15, 2023 as 10.1164/rccm.202209-1653OC

52. Swahn H, Sabith Ebron J, Lamar KM, Yin S, Kerschner JL, NandyMazumdar M*, et al.* Coordinate regulation of ELF5 and EHF at the chr11p13 CF modifier region. *J Cell Mol Med* 2019;23:7726-7740.

53. Stolzenburg LR, Yang R, Kerschner JL, Fossum S, Xu M, Hoffmann A*, et al.* Regulatory dynamics of 11p13 suggest a role for EHF in modifying CF lung disease severity. *Nucleic Acids Res* 2017;45:8773-8784.

54. Darrah RJ, Jacono FJ, Joshi N, Mitchell AL, Sattar A, Campanaro CK*, et al.* AGTR2 absence or antagonism prevents cystic fibrosis pulmonary manifestations. *J Cyst Fibros* 2019;18:127-134.

55. Gong J, Wang F, Xiao B, Panjwani N, Lin F, Keenan K*, et al.* Genetic association and transcriptome integration identify contributing genes and tissues at cystic fibrosis modifier loci. *PLoS Genet* 2019;15:e1008007.

56. Ruffin M, Mercier J, Calmel C, Mésinèle J, Bigot J, Sutanto EN*, et al.* Update on SLC6A14 in lung and gastrointestinal physiology and physiopathology: focus on cystic fibrosis. *Cell Mol Life Sci* 2020;77:3311-3323.

AJRCCM Articles in Press. Published March 15, 2023 as 10.1164/rccm.202209-1653OC

FIGURE TITLES AND LEGENDS

Figure 1. Distributions of the KNoRMA age-adjusted lung function phenotype by site cohort. Line inside each box is the median KNoRMA and the box represents the inter-quartile range (IQR), or distance between the first and third quartiles (the 25th and 75th percentiles), while violin plots show the overall population distribution. Sample sizes are shown, with symbol areas proportional to sample size. *Definitions of abbreviations*: JHU (Johns Hopkins University, twin-sibs design), UNC (University of North Carolina, extremes of phenotype), UW (University of Washington, EPIC longitudinal study for *Pseudomonas aeruginosa* effect on lung disease), FrGMS (French CF Gene Modifier Consortium, population-based) and CGS (Canadian CF Gene Modifier Study, population-based).

Figure 2. Genetic loci significantly associated with KNoRMA lung phenotype. Genome-wide Manhattan plot of associations with KNoRMA in all n = 7,840 patients. Red line shows genome-wide significance of $P < 5 \times 10^{-8}$. Blue line shows suggestive significance of $P < 5 \times 10^{-7}$.

Figure 3. Forest plots for SNP association effect size by cohort at significant loci. Beta (coefficient) refers to the average change in KNoRMA for each copy of the protective allele. Square sizes are proportional to the sample size (n) of each cohort, and the line segments are 95% confidence intervals of each beta. The most significant SNP from each locus was chosen. For each SNP, the protective allele is listed on the left and frequency of the protective allele are shown in parentheses. Cohorts are arranged by increasing mean age (at KNoRMA). *Definitions of abbreviation:* FF = patients with CF who are F508del homozygous in the *CFTR* gene; non-FF = patients with CF who are not homozygous for F508del in the *CFTR* gene.

AJRCCM Articles in Press. Published March 15, 2023 as 10.1164/rccm.202209-1653OC

Figure 4. Significant genes based on transcriptome-wide association evidence for expression versus lung function (KNoRMA). Genome-wide Manhattan plot of TWAS associations. The red line corresponds to transcriptome-wide false discovery $q < 0.10$, with significant genes labeled. Red colored text corresponds to increased expression associated with improved lung function and blue colored text corresponds to increased expression associated with decreased lung function. Regions of significant genome-wide phenotype-genotype association are marked with black arrows on the X axis.

Figure 5. Genes that drive core enrichment significant results for this branching morphogenesis pathway (GO:0048754). This VEGAS2 analysis GSEA plot includes 32 genes that are in three key signaling pathways (Shh (25); TGFb (26); and Wnt (27)) for lung development (including branching morphogenesis) and/or that interact with genes in those three signaling pathways and/or have other roles in lung development (shown in bold). The 18 genes that are associated with lung repair and/or play a role in molecular pathogenic aspects of lung disorders (e.g. COPD, asthma, lung fibrosis, cellular morphogenesis) are shown with asterisks. The remaining 11 genes are reported to have a role in development and/or morphogenesis in other tissues.

AJRCCM Articles in Press. Published March 15, 2023 as 10.1164/rccm.202209-1653OC

TABLES

Table 1. Characteristics of CF patients (all pancreatic insufficient) in this study.

Cohort	Total	CFGP WGS	Imputed	KNoRMA	Age (yrs)*	Age (yrs)	Male	European†	F508del/F508del
	n	n	n	Mean (SD)	Mean (SD)	Median	n (%)	n (%)	n (%)
JHU	1,683	1,466	217	0.54 (0.86)	20.6 (9.8)	19.0	893 (53.1)	1,565 (93.0)	947 (56.3)
UNC	2,159	1,605	554	0.60 (0.92)	26.8 (11.2)	24.9	1,170 (54.1)	2,057 (95.3)	1,606 (74.4)
UW	1,177	1,177	0	0.51 (0.73)	13.1 (3.5)	12.9	592 (50.3)	1,088 (92.4)	710 (60.3)
FrGMS	1,207	0	1,207	0.32 (0.77)	21.1 (9.2)	20.1	619 (51.3)	1,196 (99.1)	707 (58.6)
CGS	1,614	0	1,614	0.38 (0.82)	17.3 (9.2)	14.9	865 (53.6)	1,531 (94.9)	1,015 (62.9)
Overall	**7,840**	**4,248**	**3,592**	**0.48 (0.84)**	**20.6 (10.4)**	**18.4**	**4,139 (52.8)**	**7,437 (94.9)**	**4,985 (63.6)**

Definition of abbreviations: JHU (Johns Hopkins University, twin-sibs design), UNC (University of North Carolina, extremes of phenotype), UW (University of Washington, longitudinal study for effect of *Pseudomonas aeruginosa* acquisition on lung disease), FrGMS (French CF Gene Modifier Consortium, population-based) and CGS (Canadian CF Gene Modifier Study, population-based).

*Age for lung function phenotyping for KNoRMA calculation. †Based on self-reported ancestry, confirmed by ancestry-by-genotyping.

AJRCCM Articles in Press. Published March 15, 2023 as 10.1164/rccm.202209-1653OC

Table 2. Genome-wide significant ($P < 5 \times 10\text{-}8$) and suggestive ($P < 5 \times 10\text{-}7$) association results.

Chr band	Gene(s)	BP position	SNP	Risk/ Protective Allele	PAF*	Beta	*P* value	Prior GWAS regional *P* value‡
Significant findings								
3q29	*MUC20/MUC4*	195,760,866	rs2246771	G/A	0.29	0.1	6.7×10^{-12} †	3.3×10^{-11} †
5p15.33	*SLC9A3/CEP72*	537,775	rs56108664	T/C	0.83	0.11	2.8×10^{-10} †	6.8×10^{-12} †
6p21	*HLA class II*	32,462,048	rs9268860	T/C	0.68	0.08	9.9×10^{-10} †	1.2×10^{-8} †
11p13	*EHF/APIP*	34,808,842	rs485845	A/C	0.64	0.09	2.6×10^{-9} †	4.8×10^{-9} †
16p12.2	*CHP2/PRKCB*	23,779,017	rs194788	A/T	0.44	0.07	2.5×10^{-8} †	7.7×10^{-7}
Xq23	*AGTR2/SLC6A14*	116,230,240	rs12009976	G/A	0.49	0.08	6.1×10^{-12} †	1.8×10^{-9} †
Suggestive findings								
1p36	*CEP85*	26,257,354	rs41284341	A/C	0.009	0.39	1.6×10^{-7}	9.1×10^{-3}
6q15	*UBE2J1*	89,330,626	rs9294434	T/C	0.009	0.41	1.3×10^{-7}	8.0×10^{-3}
8q11.2	*SNTG1*	50,730,869	rs140650336	C/T	0.005	0.65	1.2×10^{-7}	7.2×10^{-4}
17q22	*PPM1E*	58,950,377	rs72828739	C/T	0.991	0.36	4.7×10^{-7}	7.1×10^{-2}

Gene listed if intergenic, otherwise, flanking genes are listed. *PAF=Frequency of protective allele. Beta coefficient refers to increased average KNoRMA for each copy of the protective allele. †*P* values with genome-wide significant association, $P < 5 \times 10^{-8}$; others listed are suggestive association, $P < 5 \times 10^{-7}$. ‡From Corvol et al., 2015 (6).

AJRCCM Articles in Press. Published March 15, 2023 as 10.1164/rccm.202209-1653OC

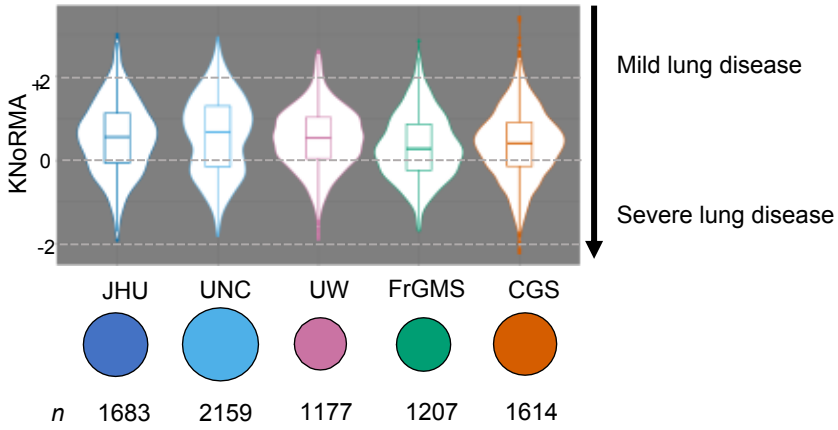


Figure 1

AJRCCM Articles in Press. Published March 15, 2023 as 10.1164/rccm.202209-1653OC

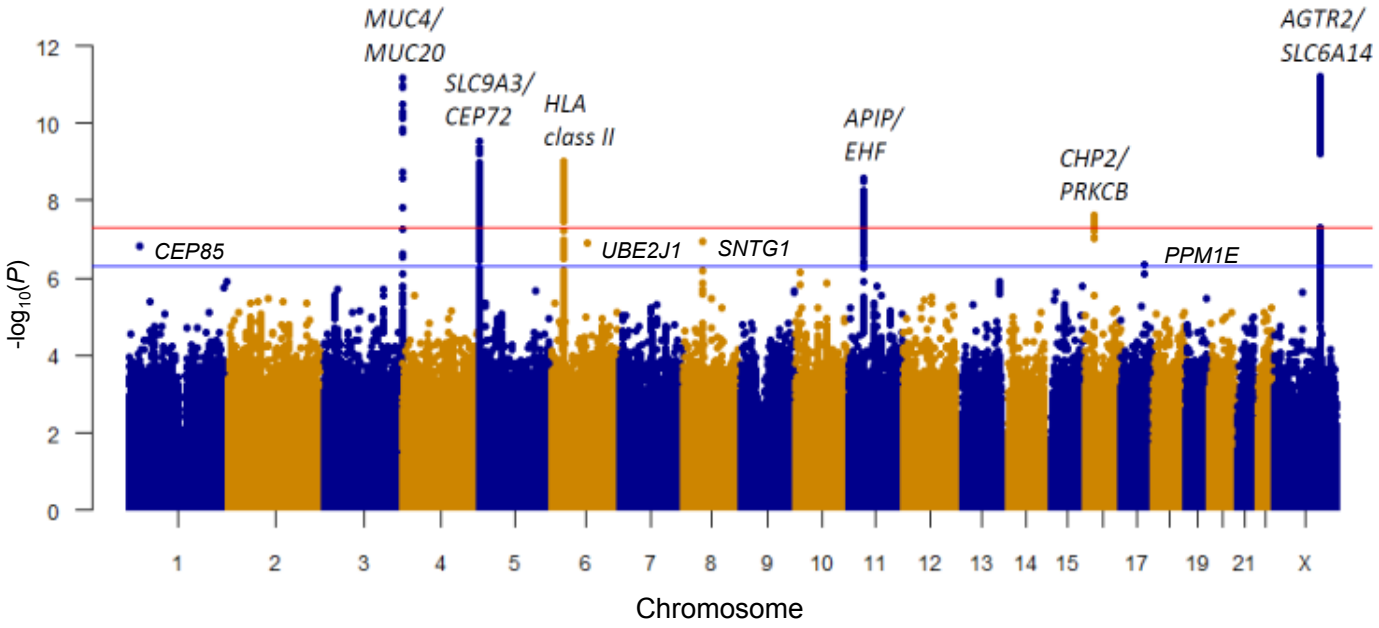


Figure 2

AJRCCM Articles in Press. Published March 15, 2023 as 10.1164/rccm.202209-1653OC

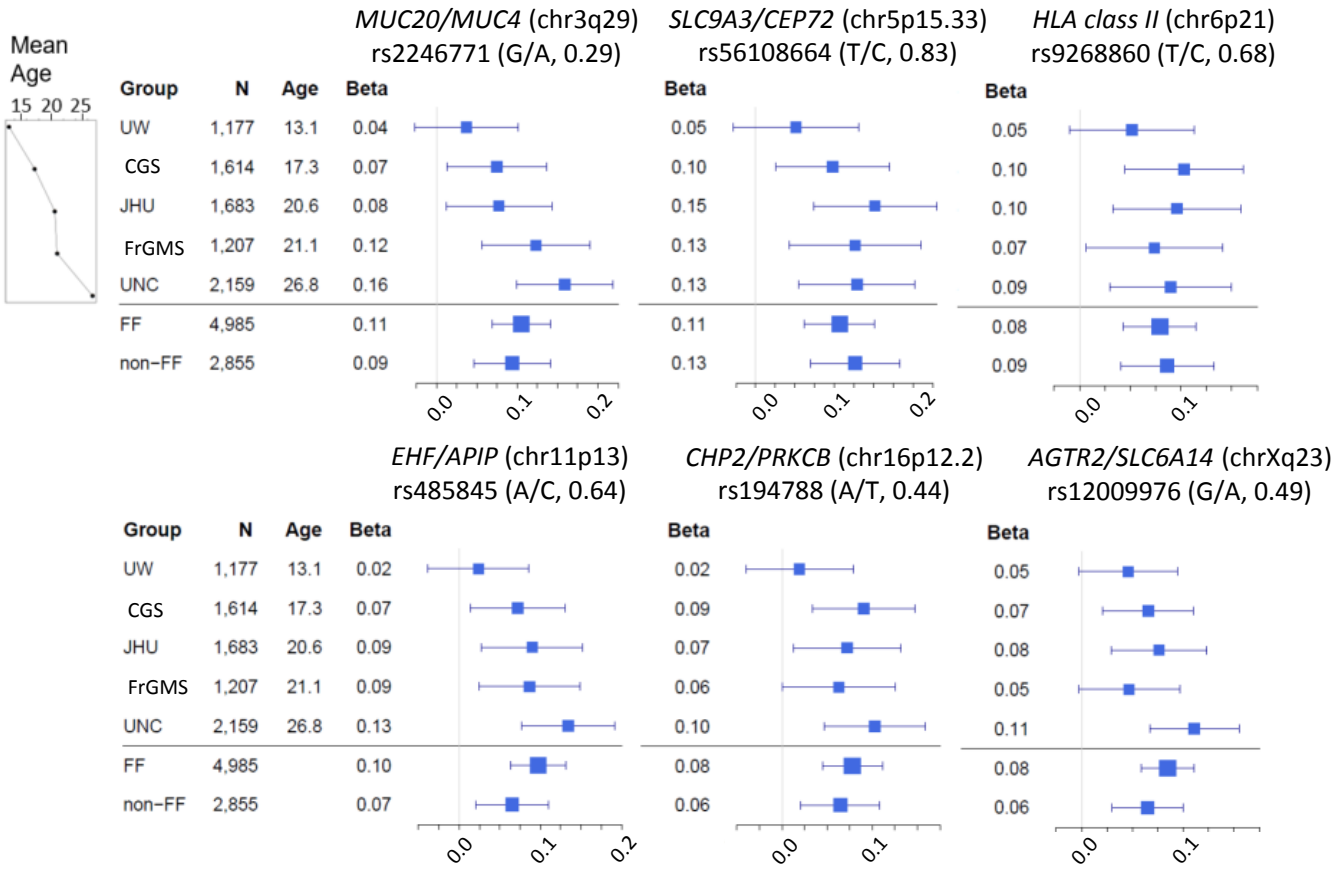


Figure 3

AJRCCM Articles in Press. Published March 15, 2023 as 10.1164/rccm.202209-1653OC

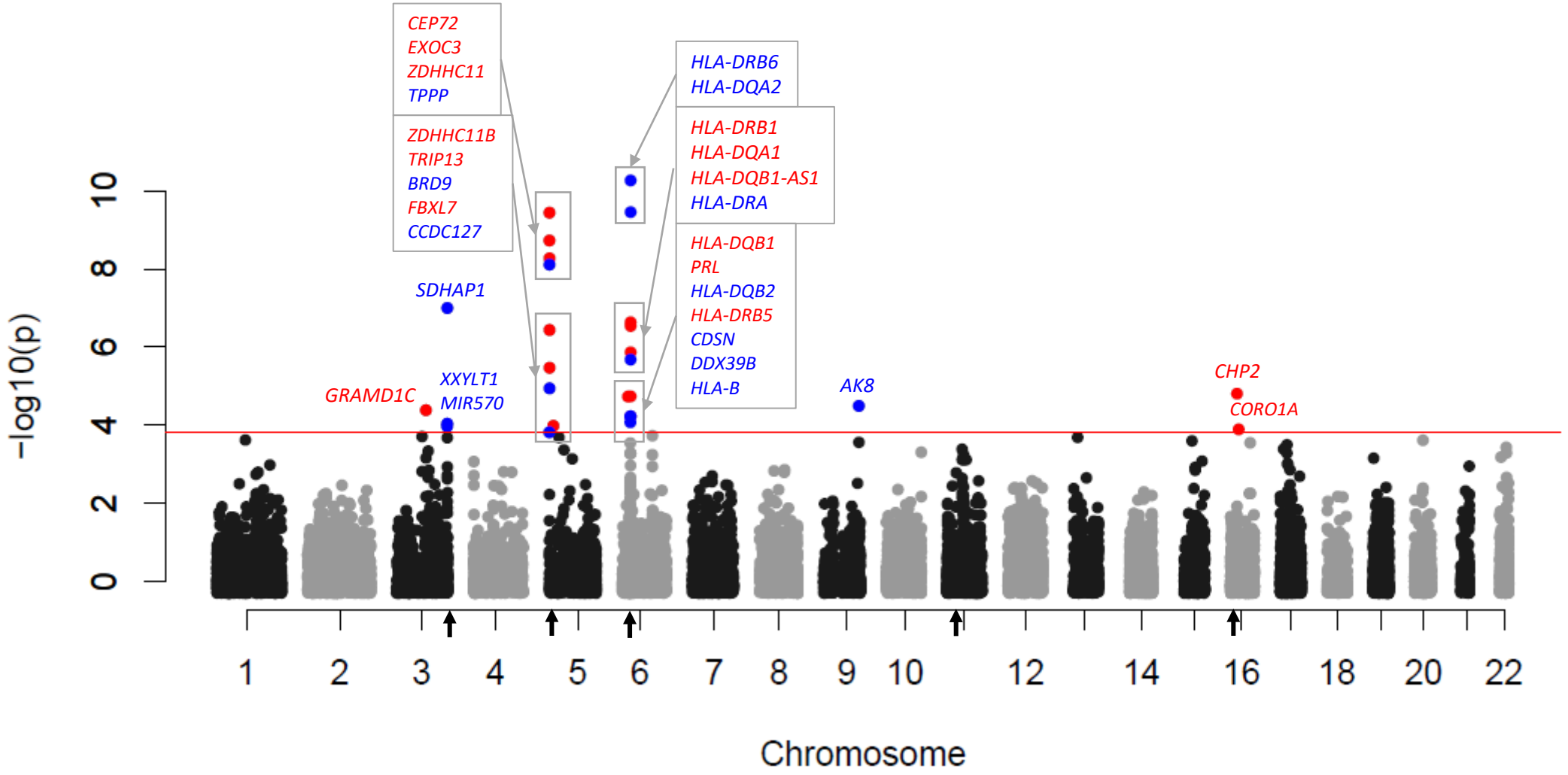
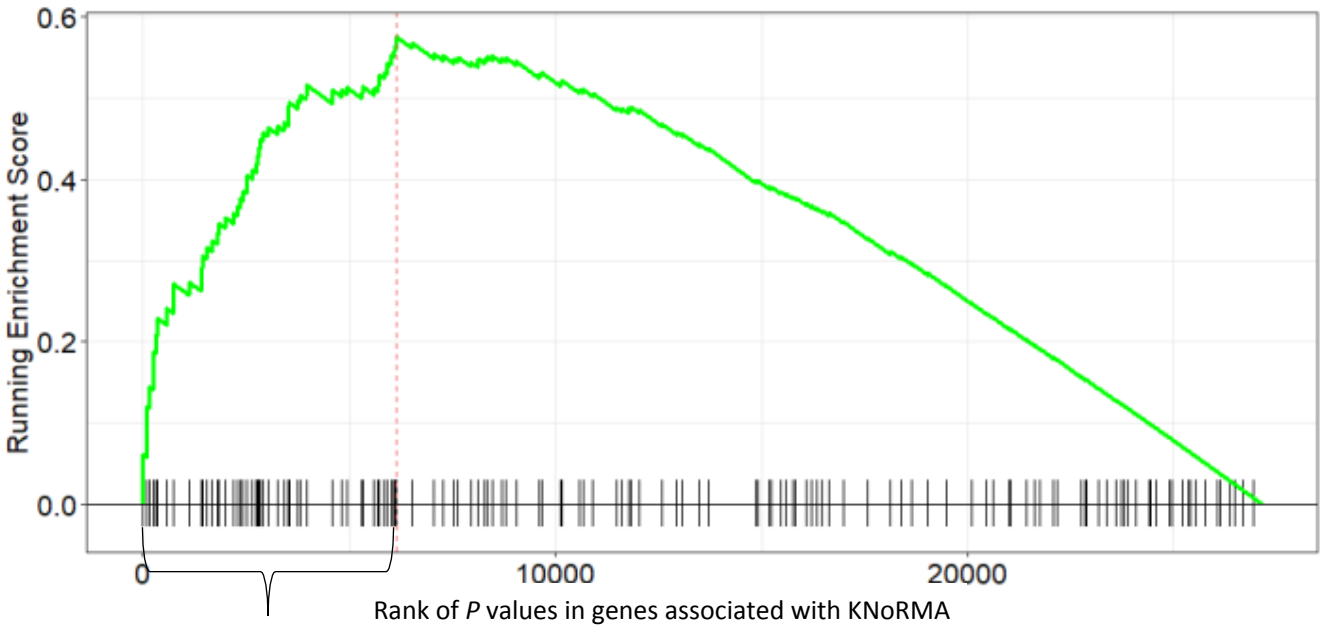


Figure 4

AJRCCM Articles in Press. Published March 15, 2023 as 10.1164/rccm.202209-1653OC



AGTR2, TNF, MET*, BMP7,* ***GLI2****,* ***PTCH1****, FGF1*,* ***GLI3****, NOTCH4*,* ***FKBPL****, MED1, NKX3-1,* ***GPC3****, CAV3*, FOXC2*, AREG*,* ***NOG****,* ***TNC****, PAK1*,* ***CTNNBIP1****, SPRY2*,* ***BMP2****,* ***SMAD4****, SIX4,* ***SOX9****,* ***WNT2****,* ***RASIP1****,* ***FOXD1****,* ***KRAS****, MIR16-1,* ***NKX2-1****,* ***PBX1****,* ***SHH****,* ***CTSH****,* ***NPNT****,* ***LAMA1****,* ***WNT2B****, PAX2*,* ***CTNNB1****,* ***SFRP2****, MAGED1*,* ***LGR4****,* ***WT1****, SEMA3E*,* ***LRP5****, HNF1B*, AR, RBM15,* ***EDNRA****,* ***VANGL2****,* ***FGF8****,* ***SMO****, VEGFA*, BCL2*,* ***WNT5A****, GDNF*, MKS1*, SIX2, HOXD11, SALL1, SPRY1**

Figure 5

AJRCCM Articles in Press. Published March 15, 2023 as 10.1164/rccm.202209-1653OC

Genetic modifiers of cystic fibrosis lung disease severity: whole genome analysis of 7,840 patients

Yi-Hui Zhou, Paul J. Gallins, Rhonda G. Pace, Hong Dang, Melis A. Aksit, Elizabeth E. Blue, Kati J. Buckingham, Joseph M. Collaco, Anna V. Faino, William W. Gordon, Kurt N. Hetrick, Hua Ling, Weifang Liu, Frankline Onchiri, Kymberleigh Pagel, Elizabeth W. Pugh, Karen S. Raraigh, Margaret Rosenfeld, Quan Sun, Jia Wen, Yun Li, Harriet Corvol, Lisa J. Strug, Michael J. Bamshad, Scott M. Blackman, Garry R. Cutting, Ronald L. Gibson, Wanda K. O'Neal, Fred A. Wright, Michael R. Knowles on behalf of the Cystic Fibrosis Genome Project

ONLINE DATA SUPPLEMENT

AJRCCM Articles in Press. Published March 15, 2023 as 10.1164/rccm.202209-1653OC

Introduction

There are many useful references about the role of different mutations in *CFTR*, nutritional aspects of disease in CF, environmental effects, impact of modifier genes, effect of modulators on lung disease severity in CF and ongoing development of new therapies in CF (1-8).

Methods

The GMS consists of unrelated patients with over-representation of extremes of lung function (6), and the TSS recruited families in which two or more surviving children have CF (9) and the CFRD study recruited pwCF both with and without CFRD (10, 11). The EPIC study is a prospective, longitudinal investigation of CF lung disease from childhood (12).

The 3,592 patients who had imputation of genome-wide genotypes were from: UNC (*n* = 554); JHU (*n* = 217); the French CF Gene Modifier Consortium (FrGMC) (*n* = 1,207); and the Canadian CF Gene Modifier Study (CGS) (*n* = 1,614).

Lung function (KNoRMA) phenotyping

KNoRMA is a validated quantitative trait of CF lung disease severity based on multiple measures of FEV_1 over 3 years, corrected for sex, age, and survival (13, 14). KNoRMA was calculated for each patient from UNC, JHU, and UW cohorts, using recent data

AJRCCM Articles in Press. Published March 15, 2023 as 10.1164/rccm.202209-1653OC

when possible, from the CF Foundation Patient Registry (CFFPR) (15) (2017), or earlier chart-extracted data. Lung function data were considered up to the last clinic visit, time of solid organ transplant, or date of death. Further, to avoid potential confounding effects of recently approved modulators to obscure genetic influences on KNoRMA, FEV_1 measures for UNC, JHU, and UW subjects were utilized only if they were obtained prior to approval of ivacaftor modulator therapy for 10 *CFTR* variants, whereby lung function measures from 2012-2014 were typically utilized. For patients with *CFTR* variants that were granted early approval for ivacaftor therapy, data from 2009-2011 were used. Finally, lung function data in encounter records in the CFFPR with mention of ivacaftor use were removed to avoid using data potentially arising from clinical trials prior to drug approval. For the FrGMC and CGS cohorts, previously derived KNoRMA values from the pre-modulator era were used (13, 14).

Whole genome sequencing

The CFGP samples were sequenced to ~30X coverage using the NovaSeq6000, performed at the Broad Institute (Cambridge, MA), using recommended QC filters as part of the GATK Best Practices Workflow, with 66 samples removed due to low sequencing coverage, high contamination, or duplication. Identity was confirmed via checks for matching sex, relatedness, and ancestry as recorded in clinical datasets, and genotypes if prior array genotypes were available, resulting in removal of 10 samples. GATK Variant Quality Score Recalibration (VQSR) was used to filter variants. The SNP VQSR model was trained using HapMap3.3 and 1KG Omni 2.5 SNP sites and a 99.6% sensitivity threshold was applied to filter variants. Specific sequencing details for the project are described previously (16). The GWAS array-based data and cohorts were

AJRCCM Articles in Press. Published March 15, 2023 as 10.1164/rccm.202209-1653OC

described previously (13), including six genotyping platforms. Genetic imputation was performed using minimac (17) and reference panels from TOPMed freeze 8 and WGS from CFGP (18). There was no sample overlap between CFGP and the imputed samples used in this study. A total of ~13.5 million bi-allelic variants were used for association testing, including ~8 million variants with minor allele frequency $\geq$ 1% and imputation Rsq ≥ 0.3, and an additional ~5.5 million variants with minor allele count $\geq$ 20 (corresponding to MAF $\geq$ 0.12%) and MAF < 1%, and imputation Rsq ≥ 0.8 (18). The focus of this study was on single-variant association testing, and all of the suggestive and significant variants are indexed by RefSNP, and thus hereafter we refer to the variants as single nucleotide polymorphisms (SNPs).

Association analysis

For each SNP, association was tested using KNoRMA as a response variable in an additive effect mixed model with covariates using the GENESIS R package (19), with z-statistics and *P* values obtained with the assocTestSingle option. Ancestry scores were computed using genotype principal components using PCAiR (20) and used as covariates in this study. As a consistency check, ancestry scores were also computed using the PCFAM (21) method, especially designed to account for family structures. The results of the two principal component methods were highly concordant, with correlation r > 0.94 for the top six principal components produced by the two methods. Analysis by the Consortium indicated that effective ancestry control for the CFGP could be achieved with as few as four genotype principal components (22), and as a conservative measure we included six principal components as covariates. Additional covariates included sex, and terms for all cohort-sites×platform combinations (as applicable). A genetic

AJRCCM Articles in Press. Published March 15, 2023 as 10.1164/rccm.202209-1653OC

relatedness matrix was used as a random effect to account for the small proportion of families as well as cryptic relatedness. As a conservative approach and to ensure no confounding due to cohort or platform, a fixed-effect meta-analysis for each of the 14 cohort-site×platform combinations weighted by sample size (23) was used to combine genotype score test *z*-statistics for SNPs with MAF $\geq$ 0.01 (for which meta-analysis was possible). For the remaining SNPs with minor allele count $\geq$ 20 and MAF < 1%, a pooled analysis was used by fitting a single mixed model with covariates including the cohort-site×platform effects. QQ plots indicated proper false positive control (**Figure E1**), and for the genome scans in all (7,840) pwCF and 4,985 F508del homozygotes, the genomic control λ (24) was 1.022 and 1.029, respectively. Results were combined using R v4.02, and plotted using the qqman package (25). Regional association plots were performed using LocusZoom (26), with the localzoom feature, with color scale using linkage disequilibrium r^2 estimates obtained as squared Pearson correlations among plotted SNPs using the samples of the current dataset. *P* value thresholds were applied at the level of genome-wide significance ($P < 5 \times 10^{-8}$) (27), and we considered SNPs with $P < 5 \times 10^{-7}$ to be strongly suggestive (28).

For each genome-wide significant region, additional analyses were conducted to search surrounding SNPs for evidence of multilocus effects for a two-locus model. To ensure that a global maximum for two-SNP models was achieved, linear regression was performed as a screening step, using the lung phenotype as a response, with all the covariates including terms for all cohort-site×platform combination, and exhaustive interrogation of all pairs of SNPs in the region to maximize the model R^2. For the SNP pair achieving maximum R^2, the SNP producing higher partial R^2 was designated as

AJRCCM Articles in Press. Published March 15, 2023 as 10.1164/rccm.202209-1653OC

primary, and the other as secondary. Finally, *P* values for the remaining SNPs were computed after conditioning on the primary SNP by including it as a fixed effect using the GENESIS mixed effect model, and similarly after conditioning on the secondary SNP. The *P* values after conditioning on the primary SNP were subjected to regional multiple correction by applying Benjamini-Hochberg false discovery correction using p.adjust in *R* v4.01. Overall, the approach was similar to conditioning on the most significant regional SNP in the initial genome scan, but with initial exhaustive testing it is possible to identify a more significant primary SNP. Moreover, the *P* values after conditioning on the secondary SNP can potentially be more significant than in the initial scan, due to reduced error variation and dependence on local linkage disequilibrium structure.

For regions with two significant SNPs, haplotype analysis was performed as a potential alternative to multi-SNP genotype modeling. The haplo.stats package (R v 4.01) was used to fit a normal likelihood ratio model (haplo.score function) for association of the pair of SNPs with KNoRMA, using the same covariates as in the main modeling, assuming a null model with no SNPs. This approach was used as an approximate screen to test for possible linkage phase effects, as it does not consider random effects due to cryptic and familial relatedness.

A reverse regression alternative to the primary single-variant analyses

As stated earlier, the KNoRMA lung function measure is designed to account for age cohort survival effects in a manner that does not reduce the power for genetic mapping. Indeed, by simply adding age as an additional predictor in the primary model above, the evidence for association drops considerably (not shown – all of the main loci effects

AJRCCM Articles in Press. Published March 15, 2023 as 10.1164/rccm.202209-1653OC

become less significant, and only the loci at 3q29, 11p13, 16p12.2, and Xq23 remain genome-wide significant). To further investigate the effect of the genotype on lung function, while considering the effect of age at phenotyping, we considered the following nested models,

Larger model: genotype~KNoRMA+ KNoRMA + Age+ KNoRMA*Age+Sex +

Cohort+genotype_PCs

Smaller model: genotype ~ Sex + Cohort+genotype_PCs

where the predictors and KNoRMA are the same as used in the primary model. As we envision genotype effects as “causal” for phenotype, this modeling may be viewed as a form of reverse regression that considers selection effects on phenotype, and (importantly) on age, which is associated with genotype by the survival of pwCF. The models were fitted using an additive logistic model in *R* for the presence of each minor allele for the genotype response. By including KNoRMA, age, and the KNoRMA*age interaction, the larger model accounts for the phenotype, as well as “genotype effects on age” that might not have been captured in the primary model. The test statistic is a likelihood ratio for the larger model vs. the smaller model, with three degrees of freedom for the accompanying chisquare statistic.

P values from this reverse regression modeling were comparable to (data not shown, within a factor of 2) the primary analysis, but generally less significant. We conclude that the reduced significance may result from the fact that a three degree of freedom test statistic is required. In addition, incorporating cryptic and familial information within this approach is not straightforward. However, the close correspondence of this reverse

AJRCCM Articles in Press. Published March 15, 2023 as 10.1164/rccm.202209-1653OC

regression approach to our primary analysis leads us to conclude that the primary analysis retains the bulk of the information content for genotype-phenotype mapping.

Expression imputation and transcriptome-wide association

Transcriptome-wide association evidence was determined using our summary association z-statistics and the published method (29). This approach uses SNP-level gene expression weights from 48 tissues from the Genotype-Tissue Expression (GTEx) project v8 (30), peripheral blood from the Netherlands Twin Register (NTR) (31), and whole blood from the Young Finns Study (YFS) (32), providing SNP-level association statistics for each gene and each of the 50 tissues. A Bonferroni correction across the tissues was performed for each gene to provide a conservative corrected p_{min} to highlight the most significant tissue. We also computed a directional omnibus statistic to aggregate signal across the tissues, as previously described (33). The approach seeks to improve upon power from other omnibus statistics by computing for each gene $z = \sum_i z_i/\sqrt{v}$, where $v = \sum_i \sum_j \rho_{ij}$ and ρ_{ij} is the observed correlation across the genes between z_i and z_j for pairs of tissues *I* and *j*. Finally, a conservative *P* value of *P* = 2Xmin(p_{min}, p_{omni}) was used as a final summary for each gene, and final Benjamini-Hochberg false discovery $q < 0.1$ was used to identify significant genes using the p.adjust package in R. Both p_{min} and p_{omni} provide direction of association of imputed expression with phenotype, and these directions were consistent for all significant genes. All genes reported in a published paper by Gusev et al. (29) were used to adjust for multiple comparisons, and pseudogenes and genes without available annotation were not included in plots. Among significant genes, an index of TWAS tissue-specificity was computed as the ratio p_{omni}/p_{min}, with large values indicative of tissue specificity.

AJRCCM Articles in Press. Published March 15, 2023 as 10.1164/rccm.202209-1653OC

Cohort effect size concordance

For the most significant SNP in each genomic region, cohort effect sizes (beta coefficients from the main association model) were compared across the five cohort sites for evidence of concordant ordering across cohorts for different loci. For each pair of significant SNPs, the Pearson correlation coefficient was computed across the five cohorts, and the average correlation across all pairs of loci was used as a statistic. Under the null hypothesis that sites do not have a consistent ordering, and assuming unlinked genomic regions, the effect size estimates should be independent. However, the statistical power to detect true signal varies somewhat across the cohorts, due to (i) varying sample size, and (ii) differing phenotypic variances due to design differences. To correct for this modest difference, rescaled effect sizes $\frac{\hat{\beta}}{\sigma\sqrt{n}}$ (where $\hat{\beta}$ denotes the cohort allelic effect, σ the overall phenotype standard deviation, and sample size n) were used for the comparisons. Using one million random permutations, the average Pearson correlations of the rescaled values under permutations were compared to the observed average correlation to construct a one-sided *P* value.

Our testing procedure for significantly consistently ranked cohort effects assumes a null hypothesis under which the effect coefficients are the same in each cohort. It is reasonable to consider whether the procedure may be biased by winner's curse effects (34), which can produce bias in overall effect size estimation for genomic discoveries. A careful consideration of the meta-analysis procedure used in this study shows that, while the winner's curse produces biases in effect size estimates, the biases are identical for each cohort because the discovery is based on the overall sample. Thus,

AJRCCM Articles in Press. Published March 15, 2023 as 10.1164/rccm.202209-1653OC

our resampling testing procedure, which compares ordering of effect estimates across loci, remains valid.

For illustration, we use the standard meta-analysis approach (35) for two constituent cohorts. Suppose that the true effect size is β for each of two cohorts indexed by $i = \{1,2\}$, i.e. the true effect size may be non-zero, but there is no heterogeneity of effect size across cohorts. We have standard errors $SE_i \approx c/\sqrt{n_i}$, where c is a constant that may depend on common quantities such as SNP minor allele frequency, and the per-cohort test statistic is $Z_i = \frac{\hat{\beta}}{SE_i} \dot{\sim} N(\mu_i, 1)$, where $\mu_i = \frac{\beta}{SE_i} = \frac{\beta\sqrt{n_i}}{c}$. A standard meta-analysis approach weights the coefficient estimates by the inverse variances (35), which for the test statistic corresponds to weights $w_i = \sqrt{\frac{n_i}{n_1+n_2}}$ (easily seen by plugging in the standard errors above), and meta $Z = (w_1 Z_1 + w_2 Z_2)$. It is easy to show that $Z \dot{\sim} N(\mu, 1)$ for $\mu = w_1\mu_1 + w_2\mu_2 = \frac{\beta}{c}\left(\frac{n_1+n_2}{\sqrt{n_1+n_2}}\right)$. To address the estimated cohort effect sizes conditioned on the statistical significance of the overall study, we consider the expectations $E\left(\hat{\beta}_i \middle| Z = z\right) = E(Z_i SE_i | Z = z) = \frac{c}{\sqrt{n_i}} E(Z_i | Z = z)$. All of the z-values are linear combinations of the other two, so each pair is bivariate normal, e.g. $(Z_i, Z) \sim N\left(\begin{bmatrix}\mu_i \\ \mu\end{bmatrix}, \begin{bmatrix}1 & w_i \\ w_i & 1\end{bmatrix}\right)$, where w_i can be shown to be the correlation between Z_i and Z. By the rules of conditional normal densities, we have $E(Z_i | Z = z) = \mu_i + w_i(z - \mu)$, so finally $E\left(\hat{\beta}_i \middle| Z = z\right) = E(Z_i SE_i | Z = z) = \frac{c}{\sqrt{n_i}} \frac{\beta\sqrt{n_i}}{c} + \frac{c}{\sqrt{n_i}} \sqrt{\frac{n_i}{n_1+n_2}} (z - \mu) = \beta + \frac{c}{\sqrt{n_1+n_2}} (z - \mu)$.

The importance of the result is as follows: (1) estimation of the effect sizes are biased according to the winner's curse, as is well known, and (2) the bias in both cohorts is the

AJRCCM Articles in Press. Published March 15, 2023 as 10.1164/rccm.202209-1653OC

same, regardless of the sample sizes, which follows from the fact that the final result no longer does not depend on cohort i (although the standard errors can be different). The result is true for each $Z = z$, and therefore applies to other constructions such as $E(\hat{\beta}_i | Z > z)$. Moreover, the result holds for an arbitrary number of cohorts (such as our cohort-site×platform combinations), as independent sets of cohorts can be collected into paired sets {one cohort group vs. a group consisting of all remaining cohorts}.

Pathway Analysis

We implemented the Gene Set Enrichment Analysis (GSEA) method (36), available in the clusterProfiler R package (37). The input is a vector of sorted gene-level *P* values from association with KNoRMA. GO and KEGG pathway annotations come from the org.Hs.eg.db R package with genome-wide annotation for Human, primarily based on mapping using Entrez Gene identifiers. We analyzed pathway sets from GO.BP, GO.CC, GO.MF, and KEGG. For each pathway, GSEA calculates an enrichment score (ES), identifies the “core enriched” genes which maximize the ES, and estimates the significance level using a *P* value from a permutation test. The *P* values are then adjusted for multiple testing using false discovery control. We ran GSEA on three sets of gene-level results: rare-variant aggregate tests, VEGAS2, and TWAS.

Non-rare variants, putative function

CAVIAR

Within each significant locus, we ran CAVIAR (Causal Variants Identification in Associated Regions), a statistical framework that quantifies the probability of each SNP to be causal (38). Inputs are SNP Z-scores from the association tests, and the pair-wise

AJRCCM Articles in Press. Published March 15, 2023 as 10.1164/rccm.202209-1653OC

correlations between each pair of SNPs. CAVIAR outputs the causal posterior probability for each SNP, and we identified the top SNPs adding up to 90% of the total probability.

VEP

We also ran the Ensembl Variant Effect Predictor (VEP), which determines the effect of variants on genes, transcripts, protein sequences, and regulatory regions (39). Inputs are the coordinates of SNPs from significant loci. VEP outputs the following: genes and transcripts affected by the variants; location of the variants (e.g. upstream of a transcript, in coding sequence, in non-coding RNA, in regulatory regions); consequence of variants on the protein sequence (e.g. stop gained, missense, stop lost, frameshift); SIFT and PolyPhen-2 scores for changes to protein sequence. Based on its variant consequence, we identified SNPs having an impact rating of high, moderate, or low.

Non-rare variants (SNPs), gene-level summary analysis

We combined *P* values from the SNP association tests to gene-level *P* values using the VEGAS2 method (40), available in the cpvSNP R package. SNPs within a flanking region of 20kb around the gene were considered part of the gene. VEGAS2 uses the matrix of correlation values among the SNPs in the gene, followed by simulations from the multivariate normal distribution, as an approximate permutation test. Gene annotations are available in the TxDb.Hsapiens.UCSC.hg38.knownGene R package. As VEGAS2 is based on permutation, we used 1 million permutations for each gene, followed by 10 million permutations for all genes showing $P < 10^{-4}$ in the first

AJRCCM Articles in Press. Published March 15, 2023 as 10.1164/rccm.202209-1653OC

permutation stage. Thus, the minimum *P* value possible by VEGAS2 was $P = 10^{-7}$. We ranked the gene-level *P* values from VEGAS2 as input for GSEA pathway analysis.

Analysis of rare variants

SNPs from the WGS passing QC, with a missing rate <2%, MAC <20, and in one of the six variant groupings defined by TOPMed (41) were aggregated into genes and analyzed in the CFGP subjects. We performed aggregate tests in the GENESIS R package using three methods: burden, SMMAT, and SKAT-O for the six variant groupings recommended by TOPMed (41). The burden test is more powerful when a large percentage of variants are associated, with similar-sized effects in the same direction. SKAT-O is more powerful when a small percentage of variants are associated and/or the effects have mixed directions. SMMAT is hybrid of the burden test and SKAT-O.

Results

To investigate possible effects of linkage phase on association, for each of the chr5p15 and chr11p13 regions, two-SNP haplotype models using the respective primary and secondary SNPs (rs56108664 and rs111275646 for *SLC9A3/CEP72*; rs483769 and rs1509661 for *EHF/APIP*) were fitted as described in **Methods**, using the same covariates as for the genotype-based modeling. The resulting haplotype-based *P* values for the two-SNP models vs a null model with no SNPs were similar to the analogous *P* values for genotype-based modeling (for *SLC9A3/CEP72*, haplotype $P = 2.37 \times 10^{-12}$, genotype $P = 5.01 \times 10^{-13}$, while for *EHF/APIP*, haplotype $P = 6.48 \times 10^{-15}$, genotype *P*

AJRCCM Articles in Press. Published March 15, 2023 as 10.1164/rccm.202209-1653OC

= 2.35 × 10^{-15}). We conclude that multi-SNP genotype-based modeling is a parsimonious approach to assess association evidence in these regions.

The previous TWAS studies took a maximal approach in which all genes with nominal (P < 0.01) gene expression associations were included for subsequent pathway investigations. In this current study, the larger sample size, combined with the more recent database GTEx v8 (20), enabled application of transcriptome-wide multiple test correction to identify individual genes achieving transcriptome-wide significance (false-discovery).

Previous studies had suggested that some GWAS loci associated more strongly with pwCF homozygous for *CFTR* F508del. Therefore, we explored genetic/genotype associations utilizing the subset of *CFTR* F508del homozygotes, analyzed in a smaller, but genetically more uniform, subset of patients (n = 4,985) (**Figure E8)**. Only the loci at chr3q29, chr11p13, and chrXq23 retained genome-wide significance in this cohort. However, the result for chr11p13 (P = 1.4 × 10^{-9}) was somewhat more significant in the *CFTR* F508del homozygote subset, compared to all pwCF (P = 2.5 × 10^{-8}) despite the reduced sample size. A test for interaction with *CFTR* F508del homozygote status at SNP rs7122048 was significant (P = 0.01, **Figure E9**). For the subset of *CFTR* F508del homozygotes, additional suggestive regions included chr5q11.2 (*ITGA2*), chr6q22.3 (*LAMA2*), and chr15 (*SV2B*) (see **Figures E8** and **E10**).

Discussion

AJRCCM Articles in Press. Published March 15, 2023 as 10.1164/rccm.202209-1653OC

Lung development, branching morphogenesis, and dysanapsis involve complex molecular, pathway signaling, biological, and physiological components. More extensive references are provided about these topics (42-48).

Aside from top GWAS hits, other genes/loci identified are noteworthy. The four loci with suggestive ($P < 5 \times 10^{-7}$) associations in 7,840 pwCF are particularly interesting, because: 1) the top SNP at each locus for *CEP85*, *UBE21J*, *SNTG1*, and *PPM1E* is intragenic (**Figure E3**) and has low MAF (0.005 – 0.009) (**Table 2**), and 2) each of these genes may be relevant to CF lung disease severity. CEP85 (like CEP72) is involved with microtubular function (49). UBE2J1 is a ubiquitin-conjugating enzyme involved in inflammation via LPS-mediated TNFα expression (50, 51). SNTG1 is a cytoplasmic peripheral membrane protein, which has been linked to lung function in GWAS (52). PPM1E is an AMPK phosphatase, which inhibits LPS-induced TNFα production in monocytes/macrophages (50). The three significant genes (*ADAMTS8*, *LINC01844*, and *PTTG1IP*) from the VEGAS2 gene-level analysis are novel, but currently available data do not suggest any obvious link to CF lung disease.

References

1. Singh VK, Schwarzenberg SJ. Pancreatic insufficiency in Cystic Fibrosis. *J Cyst Fibros* 2017;16 Suppl 2:S70-S78.

2. Ferec C, Cutting GR. Assessing the disease-liability of mutations in CFTR. *Cold Spring Harb Perspect Med* 2012;2:a009480.

AJRCCM Articles in Press. Published March 15, 2023 as 10.1164/rccm.202209-1653OC

3. Collaco JM, Blackman SM, McGready J, Naughton KM, Cutting GR. Quantification of the relative contribution of environmental and genetic factors to variation in cystic fibrosis lung function. *J Pediatr* 2010;157:802-807. e803.

4. Szczesniak R, Rice JL, Brokamp C, Ryan P, Pestian T, Ni Y*, et al.* Influences of environmental exposures on individuals living with cystic fibrosis. *Expert Rev Respir Med* 2020;14:737-748.

5. Clancy JP, Cotton CU, Donaldson SH, Solomon GM, VanDevanter DR, Boyle MP*, et al.* CFTR modulator theratyping: Current status, gaps and future directions. *J Cyst Fibros* 2019;18:22-34.

6. Drumm ML, Konstan MW, Schluchter MD, Handler A, Pace R, Zou F*, et al.* Genetic modifiers of lung disease in cystic fibrosis. *N Engl J Med* 2005;353:1443-1453.

7. Polineni D, Dang H, Gallins PJ, Jones LC, Pace RG, Stonebraker JR*, et al.* Airway Mucosal Host Defense Is Key to Genomic Regulation of Cystic Fibrosis Lung Disease Severity. *Am J Respir Crit Care Med* 2018;197:79-93.

8. Gong J, He G, Wang C, Bartlett C, Panjwani N, Mastromatteo S*, et al.* Genetic evidence supports the development of SLC26A9 targeting therapies for the treatment of lung disease. *NPJ Genom Med* 2022;7:28.

9. Vanscoy LL, Blackman SM, Collaco JM, Bowers A, Lai T, Naughton K*, et al.* Heritability of lung disease severity in cystic fibrosis. *Am J Respir Crit Care Med* 2007;175:1036-1043.

10. Aksit MA, Pace RG, Vecchio-Pagán B, Ling H, Rommens JM, Boelle PY*, et al.* Genetic modifiers of cystic fibrosis-related diabetes have extensive overlap with type 2 diabetes and related traits. *J Clin Endocrinol Metab* 2020;105:1401-1415.

AJRCCM Articles in Press. Published March 15, 2023 as 10.1164/rccm.202209-1653OC

11. Blackman SM, Commander CW, Watson C, Arcara KM, Strug LJ, Stonebraker JR*, et al.* Genetic modifiers of cystic fibrosis-related diabetes. *Diabetes* 2013;62:3627-3635.

12. Treggiari MM, Rosenfeld M, Mayer-Hamblett N, Retsch-Bogart G, Gibson RL, Williams J*, et al.* Early anti-pseudomonal acquisition in young patients with cystic fibrosis: rationale and design of the EPIC clinical trial and observational study. *Contemp Clin Trials* 2009;30:256-268.

13. Corvol H, Blackman SM, Boelle PY, Gallins PJ, Pace RG, Stonebraker JR*, et al.* Genome-wide association meta-analysis identifies five modifier loci of lung disease severity in cystic fibrosis. *Nat Commun* 2015;6:8382.

14. Taylor C, Commander CW, Collaco JM, Strug LJ, Li W, Wright FA*, et al.* A novel lung disease phenotype adjusted for mortality attrition for cystic fibrosis genetic modifier studies. *Pediatr Pulmonol* 2011;46:857-869.

15. Knapp EA, Fink AK, Goss CH, Sewall A, Ostrenga J, Dowd C*, et al.* The Cystic Fibrosis Foundation Patient Registry. Design and methods of a national observational disease registry. *Ann Am Thorac Soc* 2016;13:1173-1179.

16. Raraigh KS, Aksit MA, Hetrick K, Pace RG, Ling H, O'Neal W*, et al.* Complete CFTR gene sequencing in 5,058 individuals with cystic fibrosis informs variant-specific treatment. *J Cyst Fibros* 2021;21:463-470.

17. Howie B, Fuchsberger C, Stephens M, Marchini J, Abecasis GR. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nat Genet* 2012;44:955-959.

AJRCCM Articles in Press. Published March 15, 2023 as 10.1164/rccm.202209-1653OC

18. Sun Q, Liu W, Rosen JD, Huang L, Pace RG, Dang H*, et al.* Leveraging TOPMed imputation server and constructing a cohort-specific imputation reference panel to enhance genotype imputation among cystic fibrosis patients. *HGG Adv* 2022;3:100090.

19. Gogarten SM, Sofer T, Chen H, Yu C, Brody JA, Thornton TA*, et al.* Genetic association testing using the GENESIS R/Bioconductor package. *Bioinformatics* 2019;35:5346-5348.

20. Conomos MP, Miller MB, Thornton TA. Robust inference of population structure for ancestry prediction and correction of stratification in the presence of relatedness. *Genet Epidemiol* 2015;39:276-293.

21. Zhou YH, Marron JS, Wright FA. Computation of ancestry scores with mixed families and unrelated individuals. *Biometrics* 2018;74:155-164.

22. Kingston H, Stilp AM, Gordon W, Broome J, Gogarten SM, Ling H*, et al.* Accounting for population structure in genetic studies of cystic fibrosis. *HGG Adv* 2022;3:100117.

23. Borenstein M, Hedges LV, Higgins JPT, Rothstein HR. Introduction to meta-analysis. Hoboken: John Wiley & Sons; 2021.

24. Devlin B, Roeder K. Genomic control for association studies. *Biometrics* 1999;55:997-1004.

25. Turner SD. qqman: an R package for visualizing GWAS results using QQ and manhattan plots. *J Open Source Softw* 2018;3:731.

AJRCCM Articles in Press. Published March 15, 2023 as 10.1164/rccm.202209-1653OC

26. Boughton AP, Welch RP, Flickinger M, VandeHaar P, Taliun D, Abecasis GR*, et al.* LocusZoom.js: interactive and embeddable visualization of genetic association study results. *Bioinformatics* 2021;37:3017-3018.

27. Dudbridge F, Gusnanto A. Estimation of significance thresholds for genomewide association scans. *Genet Epidemiol* 2008;32:227-234.

28. Terzikhan N, Sun F, Verhamme FM, Adams HHH, Loth D, Bracke KR*, et al.* Heritability and genome-wide association study of diffusing capacity of the lung. *Eur Respir J* 2018;52.

29. Gusev A, Ko A, Shi H, Bhatia G, Chung W, Penninx BW*, et al.* Integrative approaches for large-scale transcriptome-wide association studies. *Nat Genet* 2016;48:245-252.

30. GTEx Consortium. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* 2020;369:1318-1330.

31. Ligthart L, van Beijsterveldt CEM, Kevenaar ST, de Zeeuw E, van Bergen E, Bruins S*, et al.* The Netherlands Twin Register: longitudinal research based on twin and twin-family designs. *Twin Res Hum Genet* 2019;22:623-636.

32. Mishra BH, Mishra PP, Raitoharju E, Marttila S, Mononen N, Sievänen H*, et al.* Modular genome-wide gene expression architecture shared by early traits of osteoporosis and atherosclerosis in the Young Finns Study. *Sci Rep* 2021;11:7111.

33. Zhou YH, Gallins PJ, Etheridge AS, Jima D, Scholl E, Wright FA*, et al.* A resource for integrated genomic analysis of the human liver. *Sci Rep* 2022;12:15151.

AJRCCM Articles in Press. Published March 15, 2023 as 10.1164/rccm.202209-1653OC

34. Ghosh A, Zou F, Wright FA. Estimating odds ratios in genome scans: an approximate conditional likelihood approach. *Am J Hum Genet* 2008;82:1064-1074.

35. Willer CJ, Li Y, Abecasis GR. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* 2010;26:2190-2191.

36. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA*, et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A* 2005;102:15545-15550.

37. Wu T, Hu E, Xu S, Chen M, Guo P, Dai Z*, et al.* clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. *Innovation (Camb)* 2021;2:100141.

38. Hormozdiari F, Kostem E, Kang EY, Pasaniuc B, Eskin E. Identifying causal variants at loci with multiple signals of association. *Genetics* 2014;198:497-508.

39. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GR, Thormann A*, et al.* The Ensembl Variant Effect Predictor. *Genome Biol* 2016;17:122.

40. Mishra A, Macgregor S. VEGAS2: Software for More Flexible Gene-Based Testing. *Twin Res Hum Genet* 2015;18:86-91.

41. Liu X, White S, Peng B, Johnson AD, Brody JA, Li AH*, et al.* WGSA: an annotation pipeline for human genome sequencing studies. *J Med Genet* 2016;53:111-112.

42. Kugler MC, Joyner AL, Loomis CA, Munger JS. Sonic hedgehog signaling in the lung. From development to disease. *Am J Respir Cell Mol Biol* 2015;52:1-13.

43. Carballo GB, Honorato JR, de Lopes GPF, Spohr T. A highlight on Sonic hedgehog pathway. *Cell Commun Signal* 2018;16:11.

AJRCCM Articles in Press. Published March 15, 2023 as 10.1164/rccm.202209-1653OC

44. Zeng LH, Barkat MQ, Syed SK, Shah S, Abbas G, Xu C*, et al.* Hedgehog Signaling: Linking Embryonic Lung Development and Asthmatic Airway Remodeling. *Cells* 2022;11:1774.

45. Aschner Y, Downey GP. Transforming Growth Factor-β: Master Regulator of the Respiratory System in Health and Disease. *Am J Respir Cell Mol Biol* 2016;54:647-655.

46. Smith BM, Hoffman EA, Rabinowitz D, Bleecker E, Christenson S, Couper D*, et al.* Comparison of spatially matched airways reveals thinner airway walls in COPD. The Multi-Ethnic Study of Atherosclerosis (MESA) COPD Study and the Subpopulations and Intermediate Outcomes in COPD Study (SPIROMICS). *Thorax* 2014;69:987-996.

47. Diaz AA, Rahaghi FN, Ross JC, Harmouche R, Tschirren J, San José Estépar R*, et al.* Understanding the contribution of native tracheobronchial structure to lung function: CT assessment of airway morphology in never smokers. *Respir Res* 2015;16:23.

48. Allen JL. Airway function throughout the lifespan: Pediatric origins of adult respiratory disease. *Pediatr Investig* 2019;3:236-244.

49. Lu B, Corey DA, Kelley TJ. Resveratrol restores intracellular transport in cystic fibrosis epithelial cells. *Am J Physiol Lung Cell Mol Physiol* 2020;318:L1145-L1157.

50. Li P, Fan JB, Gao Y, Zhang M, Zhang L, Yang N*, et al.* miR-135b-5p inhibits LPS-induced TNFα production via silencing AMPK phosphatase Ppm1e. *Oncotarget* 2016;7:77978-77986.

AJRCCM Articles in Press. Published March 15, 2023 as 10.1164/rccm.202209-1653OC

51. Polineni D, Dang H, Gallins PJ, Jones LC, Pace RG, Stonebraker JR*, et al.* Airway mucosal host defense is key to genomic regulation of cystic fibrosis lung disease severity. *Am J Respir Crit Care Med* 2018;197:79-93.

52. Wilk JB, Walter RE, Laramie JM, Gottlieb DJ, O'Connor GT. Framingham Heart Study genome-wide association: results for pulmonary function measures. *BMC Med Genet* 2007;8 Suppl 1:S8.

AJRCCM Articles in Press. Published March 15, 2023 as 10.1164/rccm.202209-1653OC

Table E1. Regional associations with KNoRMA (regions with *P* value < 10^{-5}) in all *n* = 7,840 CF patients in this study, and in *n* = 4,985 F508del homozygotes. *Definitions of abbreviations*: chr = chromosome; band = cytoband; symbol = HGNC gene symbol for nearest gene; bp = base pair position for SNP with most significant regional association, hg19; SNP = reference sequence SNP ID; MAF = minor allele frequency; *P* value = association with KNoRMA.

All CF patients

chr	band	symbol	bp	SNP	MAF	*P* value
1	1p36	*CEP85*	26257354	rs41284341	0.009	1.6×10^{-7}
1	1p32	*GLIS1*	53686818	rs141358675	0.004	4.1×10^{-6}
1	1p22	*LRRC8C*	89655585	rs142646920	0.02	8.2×10^{-6}
1	1q32	*CHIT1/BTG2*	203301135	rs80030262	0.006	7.6×10^{-6}
1	1q43	*CHRM3*	239757579	rs192349033	0.004	1.9×10^{-6}
1	1q44	*CATSPERE*	244579587	rs187260665	0.008	1.2×10^{-6}
2	2p23	*CIB4*	26581520	rs140014205	0.002	7.8×10^{-6}
2	2p16	*VRK2*	57940209	rs4603748	0.18	4.5×10^{-6}
2	2p12	*GCFC2/LRRTM4*	76436224	rs114894742	0.01	4.0×10^{-6}
2	2p12	*LRRTM4/SUCLG1*	82659835	rs17022099	0.07	8.9×10^{-6}
2	2q11	*AFF3*	99928010	rs62147651	0.28	3.3×10^{-6}
2	2q22	*NXPH2/LRP1B*	139969218	rs552730483	0.005	4.0×10^{-6}
2	2q33	*PLCL1/SATB2*	198955575	rs186647389	0.01	4.4×10^{-6}
3	3p24	*LRRC3B/NEK10*	26826536	rs59833755	0.09	2.8×10^{-6}
3	3p22	*CLASP2/PDCD6IP*	33747747	rs112132113	0.002	2.0×10^{-6}
3	3p13	*PROK2/RYBP*	71930985	rs114601176	0.01	7.8×10^{-6}
3	3p11	*CGGBP1*	88148109	rs148330308	0.002	7.3×10^{-6}
3	3q13	*TEX55*	119154262	rs534281242	0.004	9.9×10^{-6}
3	3q24	*PLSCR5/ZIC4*	147310626	rs1500866	0.36	2.0×10^{-6}
3	3q29	*MUC20/MUC4*	195760866	rs2246771	0.29	6.7×10^{-12}
4	4p15	*STIM2/PCDH7*	27882432	rs149976017	0.002	2.9×10^{-6}
4	4q26	*PDE5A/MAD2L1*	119664980	rs117249477	0.001	6.8×10^{-6}
5	5p15	*SLC9A3/CEP72*	537775	rs56108664	0.17	2.8×10^{-10}
5	5p15	*MARCHF11/ZNF622*	16402883	rs114979959	0.03	4.3×10^{-6}
5	5q11	*GPBP1/ACTBL2*	57282432	rs12654867	0.27	8.6×10^{-6}
5	5q31	*FGF1/ARHGAP26*	142698150	rs10477191	0.07	2.2×10^{-6}
6	6p24	*EEF1E1/SLC35B3*	8225758	rs187686036	0.002	4.5×10^{-6}
6	6p21	*HLA-DRA/HLA-DRB5*	32462048	rs9268860	0.32	9.9×10^{-10}
6	6q15	*UBE2J1*	89330626	rs9294434	0.009	1.3×10^{-7}
6	6q26	*PACRG*	163081553	rs34811235	0.03	7.7×10^{-6}
7	7p21	*NXPH1*	8734538	rs7799161	0.17	9.0×10^{-6}
7	7p21	*DGKB*	14244431	rs112063234	0.006	9.6×10^{-6}
7	7q21	*MAGI2/GNAI1*	79702085	rs2714649	0.48	6.1×10^{-6}
7	7q21	*GNGT1*	93648787	rs117202935	0.02	5.0×10^{-6}
8	8q11	*SNTG1*	50730869	rs140650336	0.005	1.2×10^{-7}

AJRCCM Articles in Press. Published March 15, 2023 as 10.1164/rccm.202209-1653OC

8	8q21	*SBSPON/C8orf89*	73137454	rs7819203	0.009	3.3×10^{-6}
8	8q22	*C8orf37/GDF6*	96023491	rs534774613	0.002	6.1×10^{-6}
9	9q34	*TSC1*	132915598	rs111801465	0.01	2.1×10^{-6}
10	10p14	*GATA3/CELF2*	8881405	rs72780984	0.06	8.3×10^{-6}
10	10p13	*CAMK1D*	12371661	rs184419307	0.01	6.9×10^{-7}
10	10q21	*PCDH15*	54495964	rs117393810	0.05	5.7×10^{-6}
10	10q22	*KCNMA1*	77151487	rs115644363	0.02	1.3×10^{-6}
11	11p15	*LSP1*	1886192	rs3781963	0.11	6.0×10^{-6}
11	11p13	*EHF/APIP*	34808842	rs485845	0.36	2.6×10^{-6}
11	11q12	*SCGB1D4/ASRGL1*	62321517	rs2463828	0.18	4.6×10^{-6}
11	11q13	*CHKA/KMT5B*	68149941	rs138219403	0.005	1.6×10^{-6}
11	11q14	*TENM4/FAM181B*	81116865	rs182734497	0.003	2.7×10^{-6}
11	11q22	*MMP20/MMP27*	102630776	rs2846358	0.23	7.4×10^{-6}
11	11q25	*NTM*	131968206	rs73031505	0.02	8.7×10^{-6}
12	12q13	*SLC38A4/AMIGO2*	47057702	rs148791082	0.04	5.6×10^{-6}
12	12q13	*SOAT2*	53125141	rs543477650	0.01	3.7×10^{-6}
12	12q21	*LGR5*	71463690	rs73138543	0.008	3.0×10^{-6}
12	12q24	*TBX3/MED13L*	115466937	rs80125243	0.02	5.7×10^{-6}
12	12q24	*RILPL1*	123497127	rs75279302	0.007	5.5×10^{-6}
12	12q24	*TMEM132C*	128325703	rs146366610	0.01	9.5×10^{-6}
13	13q14	*VWA8*	41583574	rs143305354	0.02	5.0×10^{-6}
13	13q33	*MYO16*	109006635	rs80032059	0.10	1.2×10^{-6}
14	14q32	*BCL11B/SETD3*	99345800	rs11160515	0.37	7.4×10^{-6}
15	15q12	*GABRG3*	27162968	rs116966042	0.02	2.4×10^{-6}
15	15q21	*AP4E1/TNFAIP8L3*	51018942	rs12324608	0.003	4.7×10^{-6}
15	15q26	*SV2B*	91293590	rs540806270	0.002	8.9×10^{-6}
15	15q26	*MCTP2/NR2F2*	95182677	rs565944044	0.003	1.6×10^{-6}
16	16p13	*PDPK1/KCTD5*	2664115	rs546441131	0.007	9.5×10^{-6}
16	16p12	*CHP2/PRKCB*	23779017	rs194788	0.44	2.5×10^{-8}
16	16q22	*ZFHX3*	73445752	rs190272774	0.007	7.7×10^{-6}
16	16q23	*WWOX*	78661539	rs117592608	0.002	6.4×10^{-6}
17	17q22	*CA10/KIF2B*	53664017	rs189245993	0.01	5.4×10^{-6}
17	17q22	*PPM1E*	58950377	rs72828739	0.009	4.7×10^{-7}
18	18p11	*L3MBTL4/ARHGAP28*	6489587	rs188115066	0.01	8.6×10^{-6}
18	18q22	*SOCS6/CBLN2*	70927742	rs75780575	0.21	9.0×10^{-6}
18	18q23	*GALR1/SALL3*	77920162	rs117399154	0.03	7.5×10^{-6}
19	19q13	*CLEC11A/GPR32*	50766899	rs138603078	0.02	3.5×10^{-6}
20	20q13	*CBLN4/MC3R*	56113464	rs559568103	0.003	8.1×10^{-6}
22	22q11	*KIAA1671*	25067863	rs187323966	0.006	7.9×10^{-6}
22	22q13	*PNPLA3/SAMM50*	43957603	rs549395716	0.003	5.8×10^{-6}
X	Xq13	*NHSL2*	72015085	rs139351147	0.005	2.3×10^{-6}
X	Xq23	*AGTR2/SLC6A14*	116230240	rs12009976	0.49	6.1×10^{-12}

AJRCCM Articles in Press. Published March 15, 2023 as 10.1164/rccm.202209-1653OC

F508del homozygote CF patients

chr	band	symbol	bp	SNP	MAF	*P* value
1	1q31	*PLA2G4A/BRINP3*	187397134	rs6425090	0.26	6.2×10^{-6}
1	1q31	*CDC73/KCNT2*	195055622	rs112526599	0.14	4.0×10^{-6}
1	1q32	*CHIT1/BTG2*	203301135	rs80030262	0.005	6.1×10^{-6}
1	1q32	*CNTN2*	205048700	rs116082426	0.003	2.3×10^{-6}
1	1q42	*RHOU/RAB4A*	229162749	rs342829	0.20	3.2×10^{-6}
1	1q43	*CHRM3*	239757579	rs192349033	0.004	9.4×10^{-6}
2	2q21	*TUBA3E/CCDC115*	130287762	rs566393574	0.004	9.9×10^{-6}
2	2q23	*ARL6IP6/RPRM*	153070652	rs1435024	0.15	7.3×10^{-6}
2	2q34	*MYL1*	210312747	rs72998411	0.009	1.4×10^{-6}
3	3p22	*TRANK1*	36881404	rs575293467	0.003	1.9×10^{-6}
3	3p22	*ULK4*	41438673	rs190729225	0.02	5.9×10^{-6}
3	3p12	*ROBO1/GBE1*	80025827	rs111563552	0.006	3.6×10^{-6}
3	3p11	*CGGBP1/ZNF654*	88080039	rs532527341	0.002	4.1×10^{-6}
3	3q29	*MUC20/MUC4*	195760866	rs2246771	0.29	4.0×10^{-8}
4	4p14	*PCDH7/ARAP2*	35904806	rs61452043	0.01	4.6×10^{-6}
5	5p15	*SLC9A3/CEP72*	582882	rs72703051	0.15	8.6×10^{-7}
5	5q11	*ITGA2*	53036259	rs2406598	0.28	1.3×10^{-7}
5	5q14	*ATP6AP1L*	82359496	rs74798522	0.02	5.3×10^{-6}
5	5q15	*FAM172A*	93896429	rs9314091	0.06	3.4×10^{-6}
5	5q21	*FBXL17/FER*	108575901	rs73217641	0.006	9.5×10^{-6}
6	6p21	*HLA-DRB1/HLA-DQA1*	32596950	rs28366340	0.39	2.8×10^{-7}
6	6q22	*LAMA2*	128927753	rs7738059	0.007	9.5×10^{-8}
6	6q23	*AKAP7*	131264361	rs146994932	0.02	6.8×10^{-6}
6	6q23	*SLC2A12*	134026173	rs75395918	0.002	9.5×10^{-6}
6	6q23	*MAP7*	136367391	rs138017191	0.004	7.8×10^{-6}
7	7p21	*DGKB*	14244431	rs112063234	0.004	5.0×10^{-6}
7	7q21	*CDK6/SAMD9*	92872390	rs10429198	0.05	8.7×10^{-6}
8	8p22	*TRMT9B*	12967580	rs144830546	0.03	6.2×10^{-6}
8	8p11	*ADAM18*	39661410	rs184059380	0.01	6.8×10^{-6}
8	8p11	*CHRNB3*	42716976	rs76252105	0.02	2.4×10^{-6}
8	8q11	*SNTG1*	50730869	rs140650336	0.005	5.4×10^{-6}
9	9q22	*TRIM14*	98086843	rs184222854	0.006	3.8×10^{-6}
10	10p11	*AL117339.4/ZNF33B*	38708056	rs191327171	0.02	6.6×10^{-6}
10	10q11	*AL117339.4/ZNF33B*	41736315	rs1608199	0.01	9.7×10^{-6}
11	11p15	*RNH1/HRAS*	521966	rs78465464	0.04	6.5×10^{-6}
11	11p13	*EHF/APIP*	34828026	rs483769	0.40	1.4×10^{-9}
11	11q22	*ARHGAP42*	100685745	rs528552437	0.004	9.3×10^{-6}
11	11q22	*MMP20/MMP27*	102630776	rs2846358	0.22	1.2×10^{-6}
12	12q13	*SLC38A4/AMIGO2*	46942803	rs79076238	0.02	1.9×10^{-6}
12	12q21	*ATP2B1/CCER1*	90074692	rs538161945	0.006	5.8×10^{-7}
13	13q12	*FGF9/SGCG*	22306992	rs9550839	0.24	6.7×10^{-6}
13	13q21	*KLHL1/DACH1*	71175920	rs141112878	0.01	4.6×10^{-6}

AJRCCM Articles in Press. Published March 15, 2023 as 10.1164/rccm.202209-1653OC

14	14q32	*INF2*	104695717	rs4326984	0.48	7.2×10^{-6}
15	15q12	*GABRG3*	27162968	rs116966042	0.02	2.6×10^{-6}
15	15q26	*SV2B/SLCO3A1*	91428896	rs4932491	0.35	2.3×10^{-7}
16	16p12	*CHP2/PRKCB*	23779286	rs2520012	0.44	2.5×10^{-6}
16	16q24	*ZCCHC14/JPH3*	87517173	rs574712649	0.004	3.8×10^{-6}
17	17p13	*GSG1L2*	9808600	rs73976614	0.003	7.7×10^{-6}
17	17p12	*HS3ST3B1/PMP22*	15096298	rs75154553	0.04	3.7×10^{-6}
17	17p11	*KCNJ18/MTRNR2L1*	22229652	rs143475159	0.07	9.0×10^{-6}
17	17q11	*BLMH*	30276810	rs56881390	0.09	4.8×10^{-6}
17	17q22	*NOG/C17orf67*	56603918	rs12603025	0.37	5.5×10^{-6}
17	17q24	*KCNJ2/SOX9*	71155392	rs9916274	0.11	7.4×10^{-6}
18	18q12	*TRAPPC8/RNF125*	31973637	rs190591278	0.006	3.7×10^{-6}
19	19q13	*PRMT1*	49678324	rs183780282	0.005	7.0×10^{-6}
20	20p11	*GGTLC1/SYNDIG1*	24152860	rs557143039	0.002	8.3×10^{-6}
21	21q22	*SUMO3/PTTG1IP*	44841475	rs235260	0.18	5.5×10^{-6}
X	Xq23	*AGTR2/SLC6A14*	116230240	rs12009976	0.49	3.0×10^{-10}

AJRCCM Articles in Press. Published March 15, 2023 as 10.1164/rccm.202209-1653OC

Table E2. CAVIAR and VEP results.

Please refer to Excel file uploaded to the journal online submission website.

AJRCCM Articles in Press. Published March 15, 2023 as 10.1164/rccm.202209-1653OC

Table E3. Transcriptome-wide association evidence for expression vs. lung function for individual annotated genes transcriptome-wide false discovery $q < 0.10$ (the table for all genes is in **Table E4**).

Column headings are as follows. (i) *Gene symbol*: HGNC gene symbol; (ii) *CHR*: chromosome; (iii) *TSS*: transcription start site in bp using human genome build hg38; (iv) *omniz*: z-statistic for the omnibus association statistic as described in Methods, with positive sign corresponding to positive association of imputed transcription values with improved lung phenotype; (v) *omnip*: two-sided *P* value for omniz; (vi) *omniq*: false discovery *q* value for omnip, corrected for all ~26K genes; (vii) *maxtissue (number of samples)*: tissue/dataset in which the most significant TWAS association occurred (eQTL dataset sample size shown in parentheses); (viii) *maxz*: z-statistic for the most significant tissue; (ix) *minp*: *P* value for the most significant tissue; (x) *minp corrected*: Bonferroni corrected *P* value for *minp,* corrected for the number of tissues with informative data; (xi) *minq*: false discovery *q* value for minp corrected, corrected for all ~26K genes; (xii) *finalp:* final *P* value = 2*[min (omnip, minp corrected)], which is a gene-level *P* value, Bonferroni-corrected for the choice of the more significant of the two *P* values (based on either the omnibus statistic or the tissue-specific statistic); (xiii) *finalq*: final *q* value correction over all 26K genes for the finalp; (xiv) *tissue specificity score* = omnip/(minp corrected), which is large when the tissue specificity *P* value is smaller than that for the omnibus statistic.

Gene symbol	CHR	TSS	omniz	omnip	omniq	maxtissue (number of samples)	maxz	minp	minp corrected	minq	finalp	finalq	tissue specificity score
HLA-DRB6	6	32527799	-6.20	2.72E-11	7.07E-07	Minor_Salivary_Gland (85)	-5.95	2.70E-09	1.27E-07	3.66E-04	5.44E-11	1.41E-06	2.14E-04
HLA-DQA2	6	32709119	-5.94	1.76E-10	1.58E-06	Brain_Hypothalamus (108)	-5.54	3.05E-08	1.43E-06	1.96E-03	3.52E-10	3.15E-06	1.23E-04
CEP72	5	612387	5.94	1.82E-10	1.58E-06	Thyroid (399)	6.53	6.65E-11	1.53E-09	1.99E-05	3.64E-10	3.15E-06	1.19E-01
EXOC3	5	443273	5.61	1.63E-09	8.48E-06	Brain_Substantia_nigra (80)	6.69	2.28E-11	9.37E-10	1.99E-05	1.87E-09	1.22E-05	1.74E+00
ZDHHC11	5	851101	4.48	1.50E-06	3.55E-03	Adrenal_Gland (175)	6.46	1.08E-10	2.69E-09	2.33E-05	5.38E-09	2.33E-05	5.58E+02
TPPP	5	693510	-4.31	3.75E-06	7.49E-03	Brain_Cerebellum (154)	-6.35	2.16E-10	3.89E-09	2.53E-05	7.78E-09	2.89E-05	9.64E+02
SDHAP1	3	195717187	-5.07	5.03E-08	1.87E-04	Testis (225)	-5.82	5.91E-09	2.36E-07	4.72E-04	1.01E-07	2.61E-04	2.13E-01
HLA-DRB1	6	32557625	4.83	2.12E-07	6.13E-04	Small_Intestine_Terminal_Ileum (122)	5.92	3.23E-09	1.16E-07	3.66E-04	2.32E-07	5.02E-04	1.83E+00
HLA-DQA1	6	32595956	4.52	1.21E-06	3.13E-03	Small_Intestine_Terminal_Ileum (122)	5.87	4.38E-09	1.45E-07	3.76E-04	2.90E-07	5.80E-04	8.34E+00
ZDHHC11B	5	767067	3.94	2.26E-05	2.80E-02	Thyroid (399)	5.70	1.22E-08	1.83E-07	4.31E-04	3.66E-07	6.79E-04	1.23E+02
HLA-DQB1-AS1	6	32628132	4.27	4.42E-06	8.20E-03	Colon_Sigmoid (203)	5.63	1.84E-08	6.82E-07	1.18E-03	1.36E-06	2.08E-03	6.48E+00
HLA-DRA	6	32407619	-1.91	4.06E-02	8.96E-01	Adipose_Subcutaneous (385)	-5.25	1.50E-07	1.05E-06	1.61E-03	2.10E-06	2.87E-03	3.87E+04
TRIP13	5	892940	4.12	9.55E-06	1.38E-02	Artery_Aorta (267)	4.92	8.52E-07	1.70E-06	2.21E-03	3.40E-06	4.42E-03	5.62E+00
BRD9	5	892757	-4.22	5.71E-06	9.89E-03	Brain_Cerebellum (154)	-3.85	1.17E-04	8.18E-04	2.80E-01	1.14E-05	1.41E-02	6.98E-03
CHP2	16	23765948	0.99	2.86E-01	1.00E+00	Lung (383)	4.70	2.60E-06	7.81E-06	9.46E-03	1.56E-05	1.78E-02	3.66E+04
HLA-DQB1	6	32636160	4.13	9.16E-06	1.38E-02	Brain_Putamen_basal_ganglia (111)	4.10	4.21E-05	2.02E-03	4.61E-01	1.83E-05	1.85E-02	4.54E-03
PRL	6	22297730	3.50	1.70E-04	1.30E-01	Brain_Hippocampus (111)	4.58	4.64E-06	9.27E-06	1.05E-02	1.85E-05	1.85E-02	1.83E+01

AJRCCM Articles in Press. Published March 15, 2023 as 10.1164/rccm.202209-1653OC

AK8	9	135754164	-4.02	1.61E-05	2.21E-02	Brain_Substantia_nigra (80)	-4.21	2.55E-05	6.63E-04	2.69E-01	3.22E-05	3.10E-02	2.43E-02
GRAMD1C	3	113547029	3.96	2.07E-05	2.69E-02	Artery_Coronary (152)	4.07	4.70E-05	6.58E-04	2.69E-01	4.14E-05	3.84E-02	3.15E-02
HLA-DQB2	6	32731311	-3.89	2.95E-05	3.26E-02	Brain_Cerebellum (154)	-4.31	1.60E-05	7.19E-04	2.79E-01	5.90E-05	4.74E-02	4.10E-02
HLA-DRB5	6	32498064	1.99	3.28E-02	8.59E-01	NTR.BLOOD.RNAARR (1247)	4.98	6.25E-07	3.00E-05	3.12E-02	6.00E-05	4.74E-02	1.09E+03
CDSN	6	31088223	-3.88	3.01E-05	3.26E-02	Skin_Sun_Exposed_Lower_leg (414)	-3.82	1.34E-04	2.68E-04	1.39E-01	6.02E-05	4.74E-02	1.12E-01
DDX39B	6	31510225	-2.24	1.59E-02	7.50E-01	Brain_Cerebellar_Hemisphere (125)	-4.53	6.04E-06	4.23E-05	3.92E-02	8.46E-05	5.94E-02	3.76E+02
MIR570	3	195426272	-3.49	1.78E-04	1.32E-01	Cells_Transformed_fibroblasts (300)	-4.79	1.70E-06	4.60E-05	4.12E-02	9.20E-05	6.29E-02	3.86E+00
FBXL7	5	15500305	0.90	3.36E-01	1.00E+00	Colon_Sigmoid (203)	4.53	5.80E-06	5.22E-05	4.52E-02	1.04E-04	6.96E-02	6.43E+03
XXYLT1	3	194991896	-0.83	3.72E-01	1.00E+00	Heart_Left_Ventricle (272)	-4.77	1.80E-06	5.39E-05	4.52E-02	1.08E-04	7.00E-02	6.91E+03
CORO1A	16	30194148	3.72	6.42E-05	6.18E-02	YFS.BLOOD.RNAARR (1264)	3.65	2.62E-04	7.85E-04	2.79E-01	1.28E-04	8.14E-02	8.18E-02
CCDC127	5	218330	-2.16	2.01E-02	7.94E-01	Brain_Nucleus_accumbens_basal_ganglia (130)	-4.73	2.24E-06	7.63E-05	6.20E-02	1.53E-04	9.44E-02	2.63E+02

AJRCCM Articles in Press. Published March 15, 2023 as 10.1164/rccm.202209-1653OC

Table E4. TWAS results for all genes (n = 25,982) using the CFGP WGS for 7,840 patients.

Please refer to Excel file uploaded to the journal online submission website.

AJRCCM Articles in Press. Published March 15, 2023 as 10.1164/rccm.202209-1653OC

Table E5. Association results of 36,946 genes and Consortium lung phenotype (KNoRMA).

Please refer to Excel file uploaded to the journal online submission website.

AJRCCM Articles in Press. Published March 15, 2023 as 10.1164/rccm.202209-1653OC

Table E6. Pathways that are significantly associated with Consortium lung phenotype (KNoRMA).

Please refer to Excel file uploaded to the journal online submission website.

AJRCCM Articles in Press. Published March 15, 2023 as 10.1164/rccm.202209-1653OC

Supplementary Figures

AJRCCM Articles in Press. Published March 15, 2023 as 10.1164/rccm.202209-1653OC

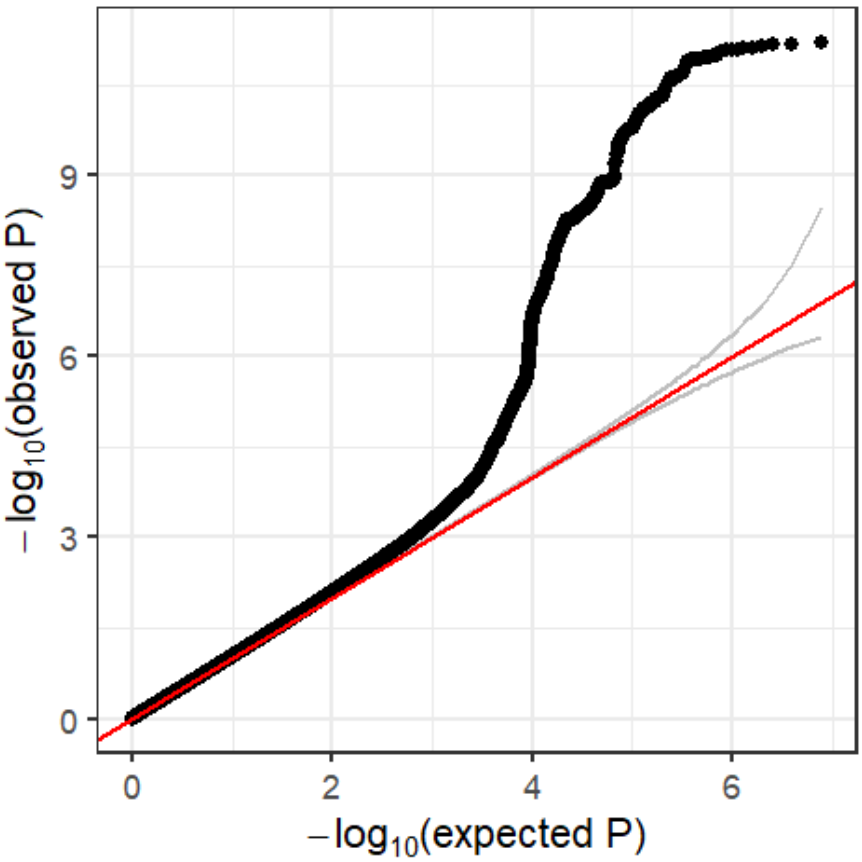
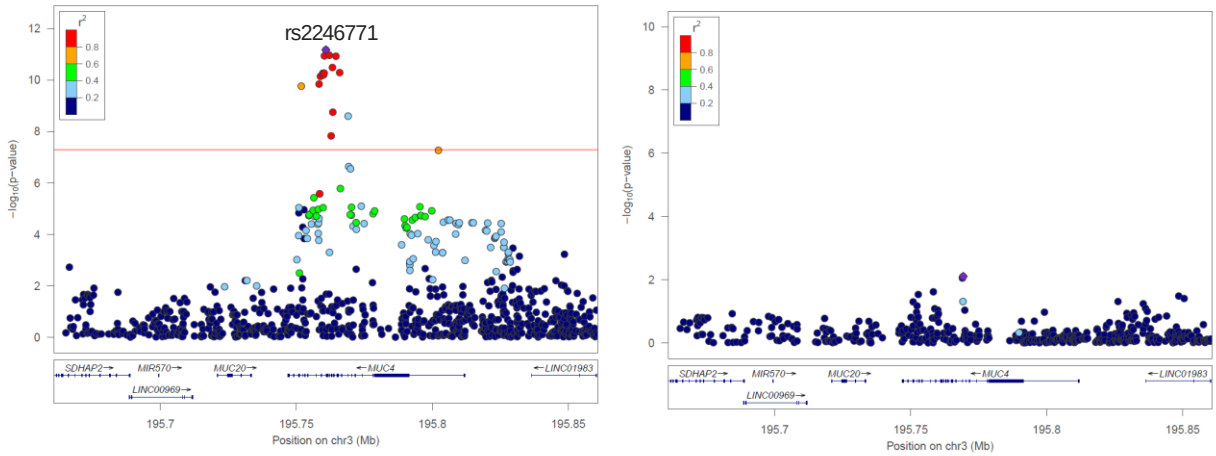


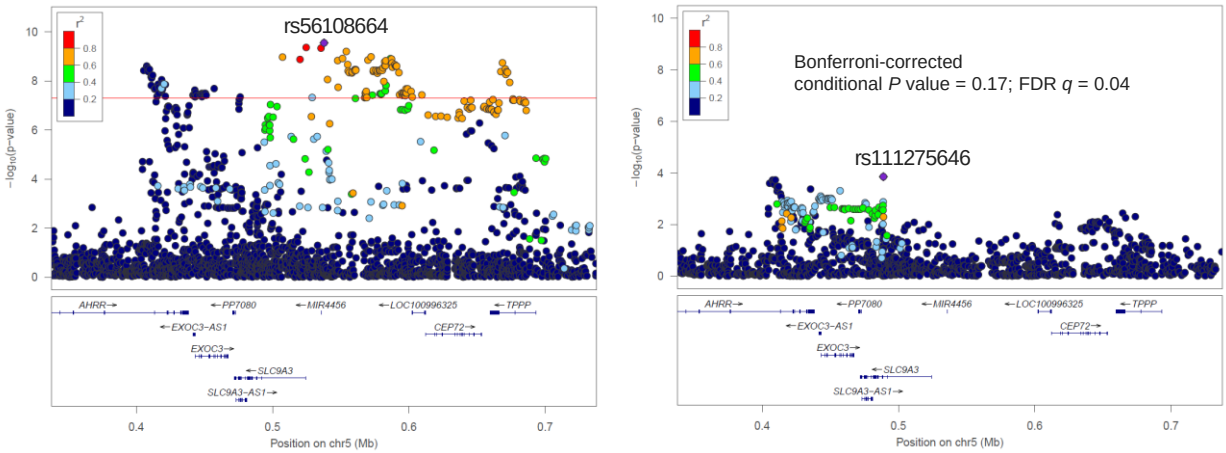
Figure E1. Control of false positives. The quantile-quantile plot of observed vs. expected *P* values shows proper control of false positives (genomic control λ = 1.022). Results shown are from genome-wide association analysis with KNoRMA in all participants (*n* = 7,840).

AJRCCM Articles in Press. Published March 15, 2023 as 10.1164/rccm.202209-1653OC

MUC4/MUC20 (chr3q29)



SLC9A3/CEP72 (chr5p15.33)



HLA class II (chr6p21)

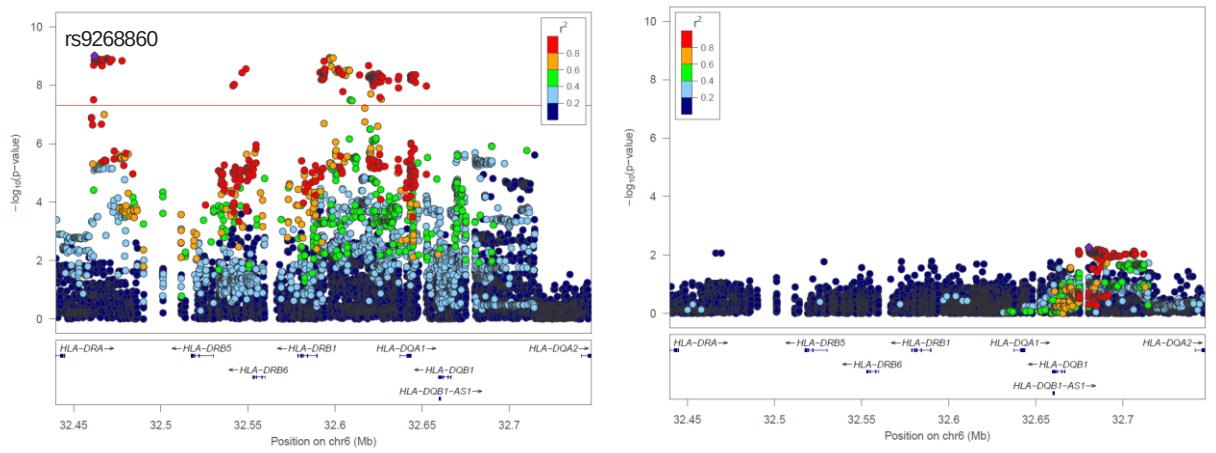
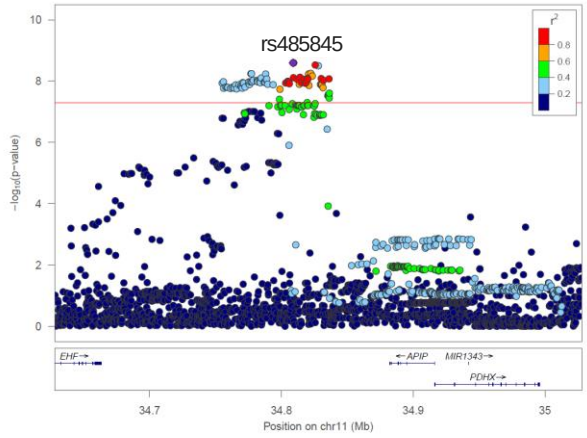
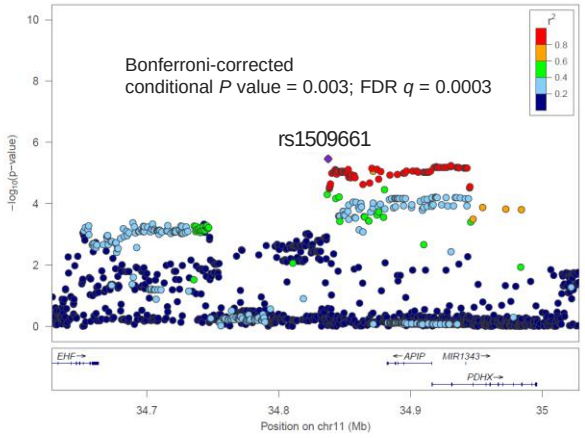


Figure E2. LocusZoom plots for six loci with genome-wide significant associations. For each locus, the left panel shows the *P* values from the original association scan, and the right panel shows the conditional *P* values after conditioning on the most significant regional SNP. Red line on left panels, $P < 5 \times 10^{-8}$ (also see **Table 2).** For two of the loci (*SLC9A3/CEP72* and *EHF/APIP*), regional correction for the secondary/conditional SNP was significant after regional false discovery rate correction.

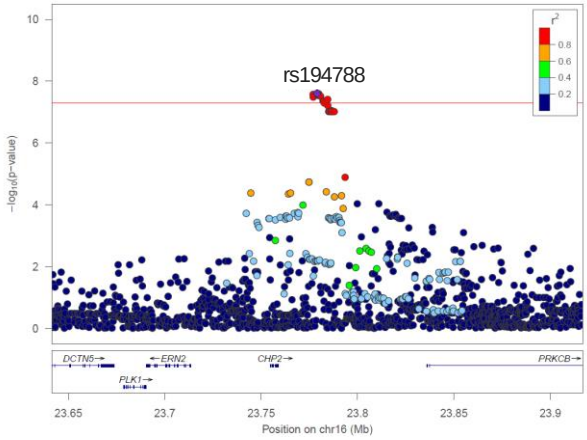
AJRCCM Articles in Press. Published March 15, 2023 as 10.1164/rccm.202209-1653OC

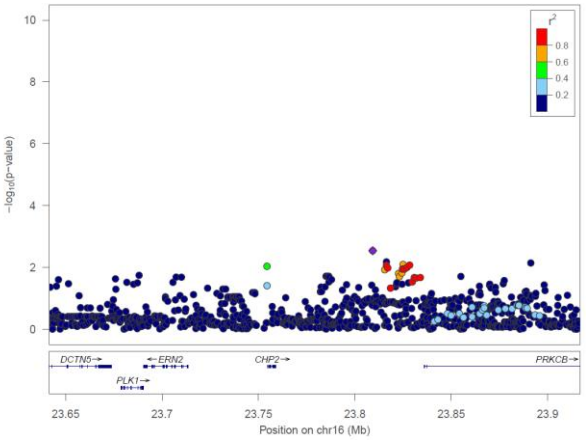
EHF/APIP (chr11p13)



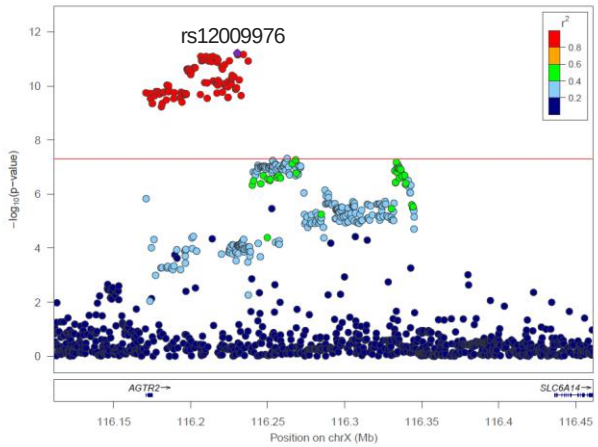


CHP2/PRKCB (chr16p12.2)





AGTR2/SLC6A14 (chrXq23)



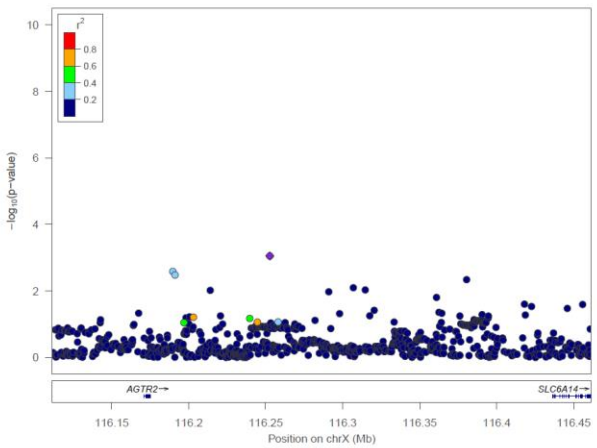
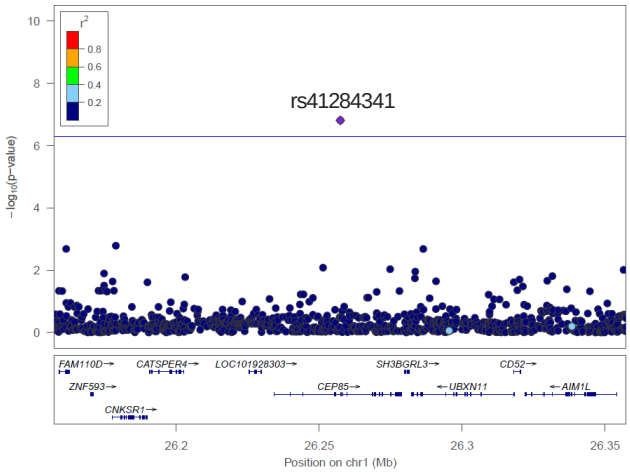


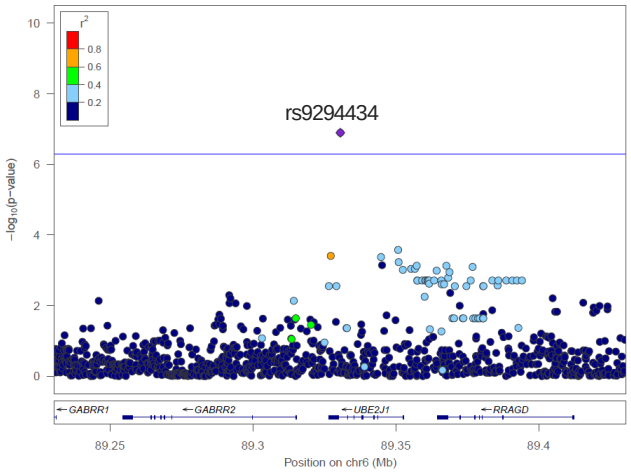
Figure E2 (continued). LocusZoom plots for six loci with genome-wide significant associations. For each locus, the left panel shows the *P* values from the original association scan, and the right panel shows the conditional *P* values after conditioning on the most significant regional SNP. Red line on left panels, $P < 5 \times 10^{-8}$ (also see **Table 2).** For two of the loci (*SLC9A3/CEP72* and *EHF/APIP*), regional correction for the secondary/conditional SNP was significant after regional false discovery rate correction.

AJRCCM Articles in Press. Published March 15, 2023 as 10.1164/rccm.202209-1653OC

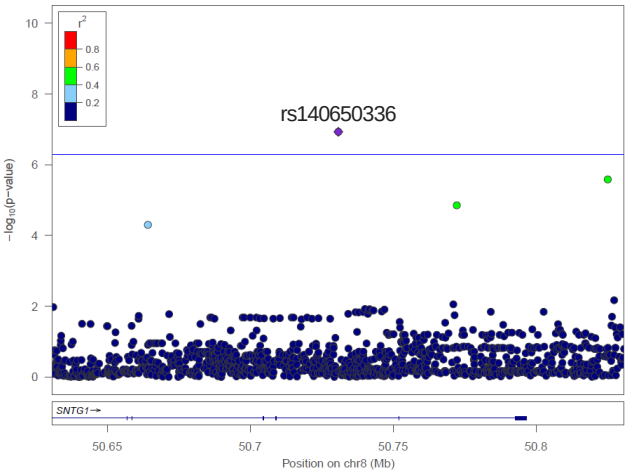
CEP85 (chr1p36)



UBE2J1 (chr6q15)



SNTG1 (chr8q11.2)



PPM1E (chr17q22)

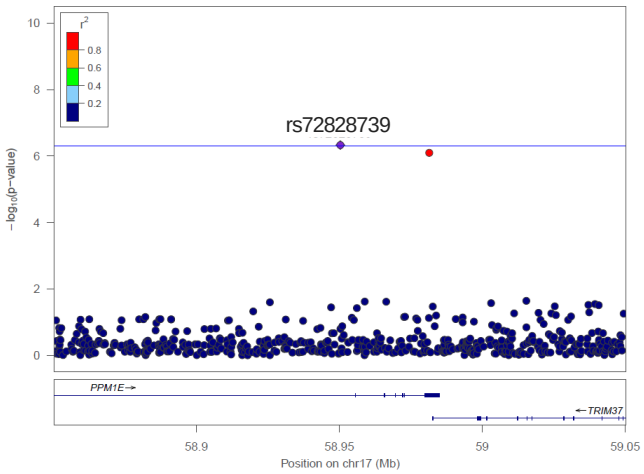


Figure E3. LocusZoom plots for four loci with suggestive associations in all patients. Blue line, $P < 5 \times 10^{-7}$ (see also **Table 2**).

AJRCCM Articles in Press. Published March 15, 2023 as 10.1164/rccm.202209-1653OC

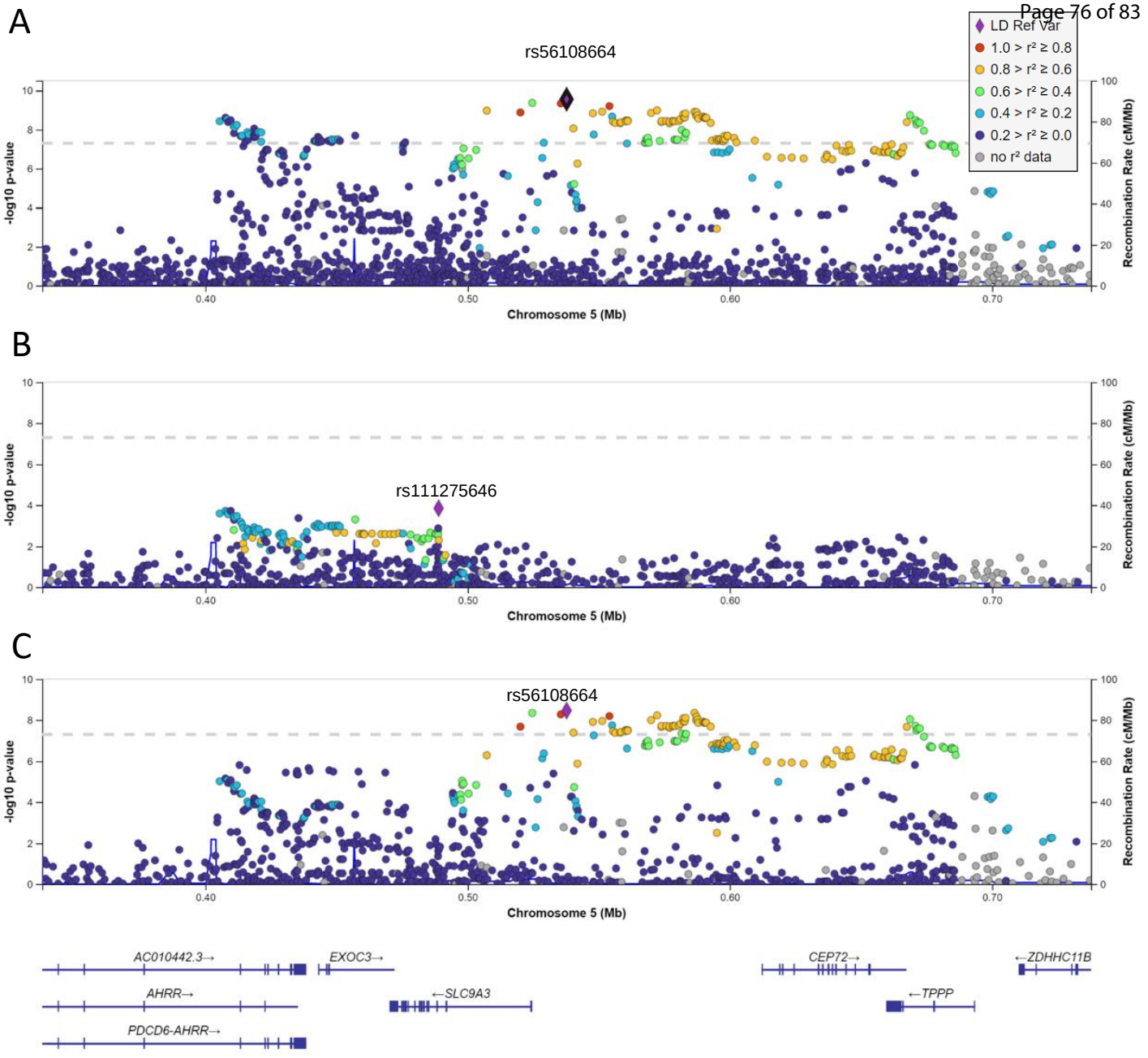


Figure E4. LocusZooms in the chr5p15.33 region showing *P* values for original and conditional analyses for the most significant SNPs. Exhaustive two-SNP modeling identified rs56108664 ("primary") and rs111275646 ("secondary") as the most significant pair of SNPs (linkage disequilibrium $r^2 = 0.01$ for rs56108664/rs111275646) in predicting the KNoRMA phenotype in the all-patients analysis. These SNPs are highlighted on the plots. Dashed line shows genome-wide significance of $P < 5 \times 10\text{-}8$. **(A)** in original genome-wide association scan, repeated here for easy comparison; **(B)** after controlling for the primary SNP rs56108664 (protective allele frequency, PAF = 0.83) determined from exhaustive two-SNP modeling in the region; **(C)** after controlling for the secondary SNP rs111275646 (PAF = 0.85). The result after conditioning on the secondary SNP is similar to that of the initial single-SNP scan.

AJRCCM Articles in Press. Published March 15, 2023 as 10.1164/rccm.202209-1653OC

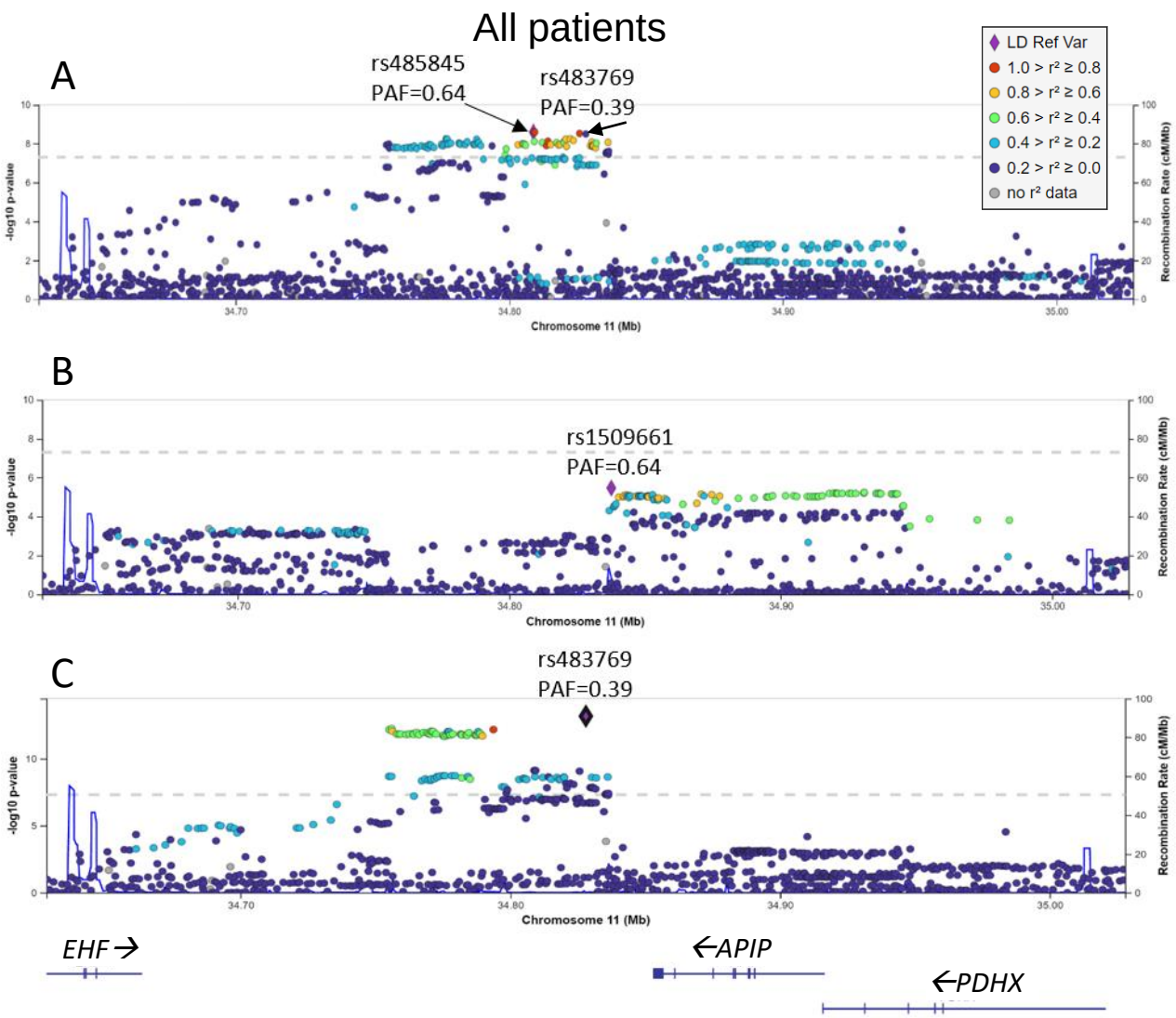
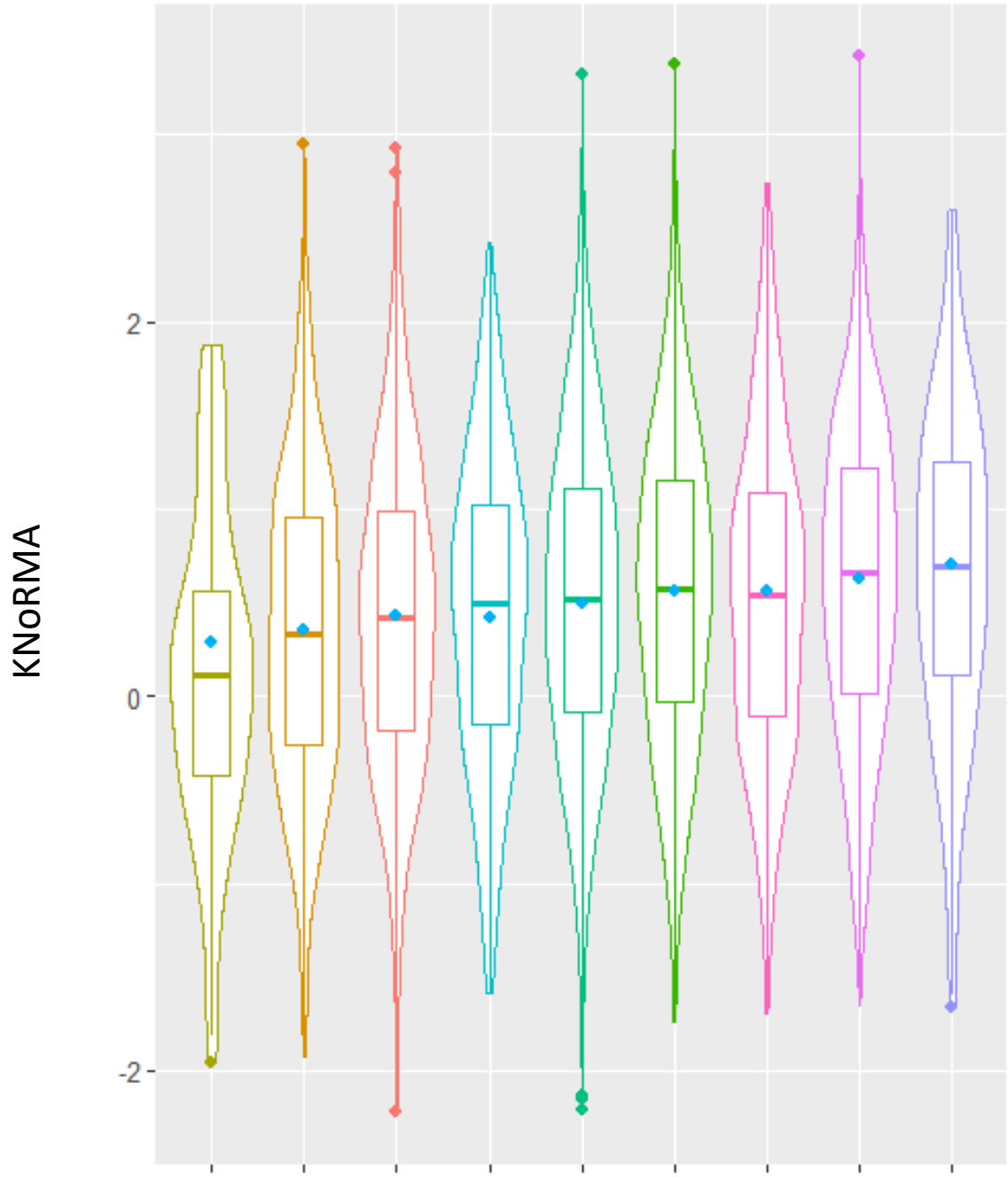


Figure E5. LocusZooms in the chr11p13 region showing *P* values for original and conditional analyses for the most significant SNPs. Exhaustive two-SNP modeling identified rs483769 ("primary") and rs1509661 ("secondary") as the most significant pair of SNPs (linkage disequilibrium $r^2 = 0.28$ for rs483769/rs1509661) in predicting the KNoRMA phenotype in the all-patients analysis. These SNPs and their protective allele frequencies (PAF) are highlighted on the plots. Dashed line shows genome-wide significance of $P < 5 \times 10^{-8}$. **(A)** Original genome-wide association scan in all patients, repeated here for easy comparison. The most significant SNP in this single-SNP analysis was rs485845 ($P = 2.6 \times 10^{-9}$), which is in the same LD block and correlated with the primary (rs483769) SNP ($r^2 = 0.35$ for rs483769/rs485845). **(B)** *P* values for all patients after controlling for the primary SNP (minimum $P = 3.4 \times 10^{-6}$ at rs1509661), with regional Bonferroni $P = 0.003$. **(C)** *P* values after controlling for the secondary SNP rs1509661 are much smaller (4-5 orders of magnitude, minimum $P = 7.2 \times 10^{-14}$) than the *P* values from the original scan.

AJRCCM Articles in Press. Published March 15, 2023 as 10.1164/rccm.202209-1653OC



rs483769	**A/A**	**A/A**	**A/A**	**A/G**	**A/G**	**A/G**	**G/G**	**G/G**	**G/G**
rs1509661	T/T	G/T	G/G	T/T	G/T	G/G	T/T	G/T	G/G
n	76	749	2,100	395	2,277	1,026	561	533	123
KNoRMA	0.12	0.34	0.43	0.45	0.52	0.56	0.53	0.62	0.65

Figure E6. Violin/boxplots of the KNoRMA phenotype for various genotype combinations for rs483769 and rs1509661 in the EHF/APIP region. The two SNPs indicated were the most significant in the primary and secondary regions based upon the two-SNP additive model (see **Figure E5**). "Risk" alleles, i.e., those that associate with lower lung function (KNoRMA), are highlighted in red. Line inside each box is the median KNoRMA and the box represents the inter-quartile range (IQR), or distance between the first and third quartiles (the 25th and 75th percentiles). Blue dots are predicted KNoRMA based on the effect sizes of each SNP in the two-SNP additive model. Violin plots represent the phenotype distribution. The number of study participants carrying each genotype combination (*n*) is shown along with the mean (KNoRMA).

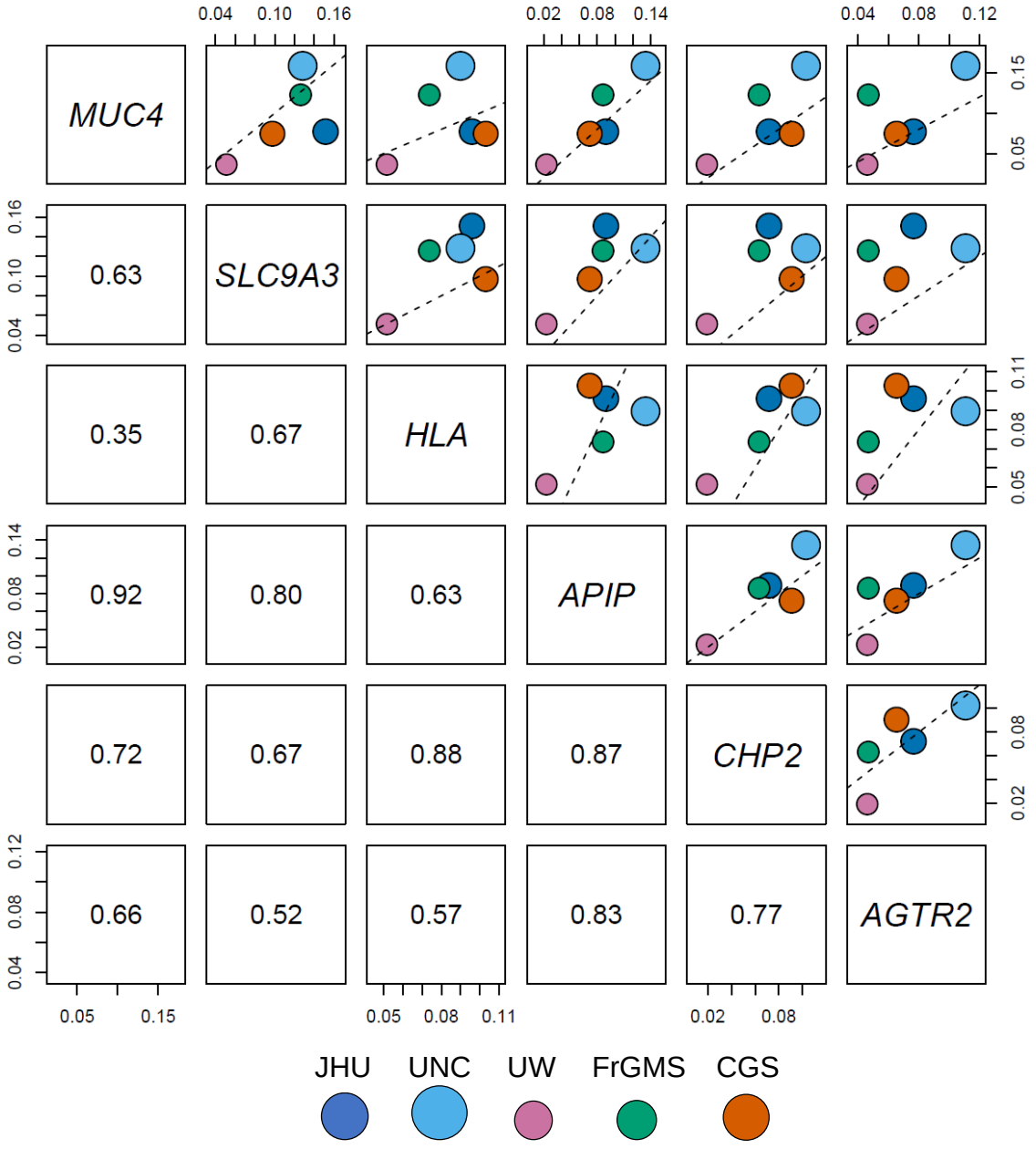


Figure E7. Consistent ordering of cohort effect sizes across multiple loci. Allelic effect sizes (beta values) from forest plots of Figure 3 are plotted as pairwise scatterplots, with bubbles corresponding to site cohort and with bubble areas proportional to cohort size. The protective allele is used as reference, so all are in the same positive direction, and the unit line is shown as a dashed line. The results show positive correlation for all pairs of cohorts across the six genome-wide significant loci, indicating consistent ordering of cohort effects (P = 0.0014) by the most conservative assumptions.

AJRCCM Articles in Press. Published March 15, 2023 as 10.1164/rccm.202209-1653OC

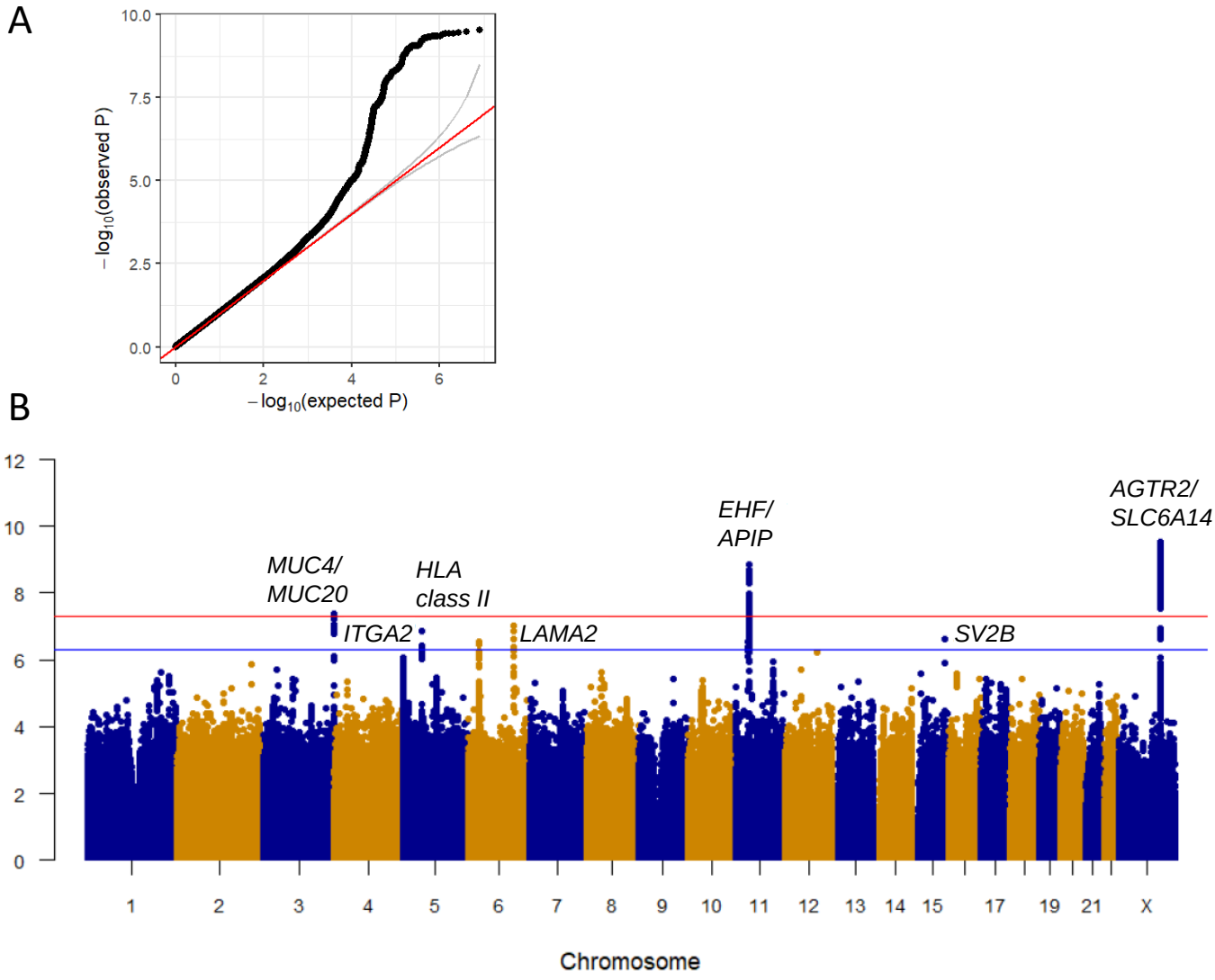


Figure E8. Results from genome-wide Manhattan plot of associations with KNoRMA in the Phe508del homozygotes (*n* = 4,985). (A) The quantile-quantile plot of observed vs. expected *P* values shows proper control of false positives (genomic control λ = 1.029). (B) Manhattan plot, with red and blue lines corresponding to significant ($P < 5 \times 10^{-8}$) and suggestive ($P < 5 \times 10^{-7}$) associations, respectively.

AJRCCM Articles in Press. Published March 15, 2023 as 10.1164/rccm.202209-1653OC

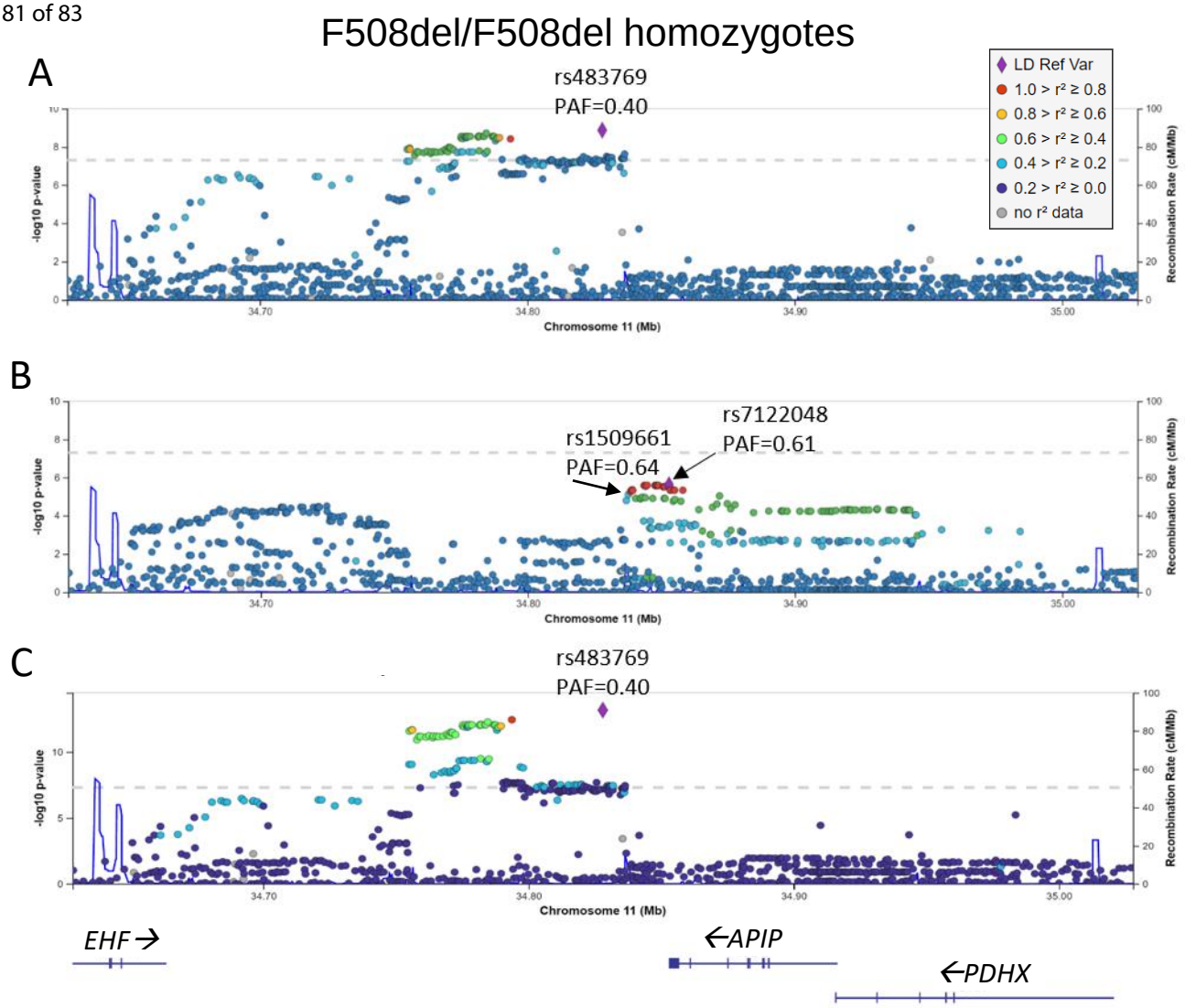
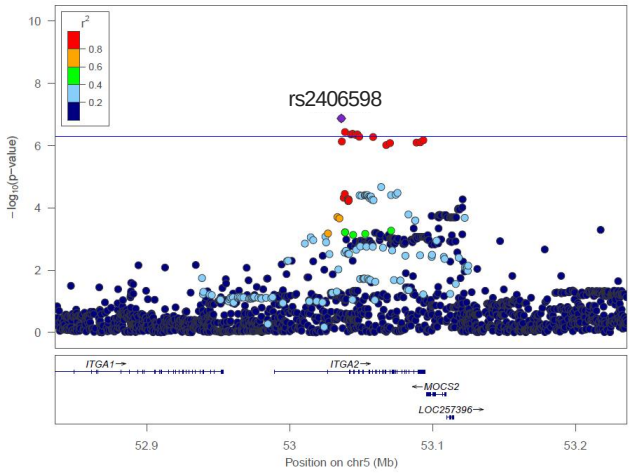


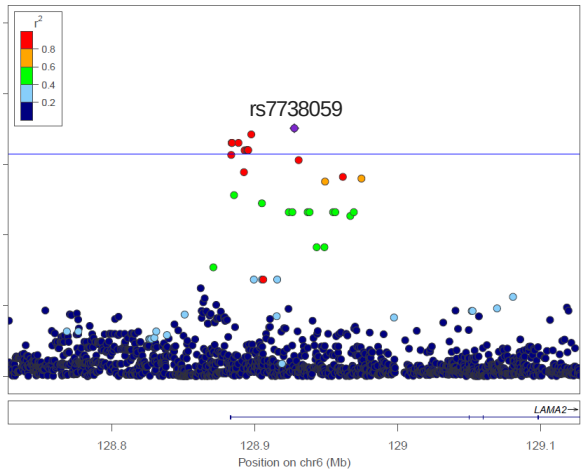
Figure E9. LocusZooms in the chr11p13 region showing *P* values for original and conditional analyses for the most significant SNPs. Exhaustive two-SNP modeling identified rs483769 ("primary") and rs1509661 ("secondary") as the most significant pair of SNPs (linkage disequilibrium $r^2 = 0.28$ for rs483769/rs1509661) in predicting the KNoRMA phenotype in the all-patients analysis. These two SNPs are used throughout as primary/secondary for consistency, and these SNPs and others and their protective allele frequencies (PAF) are highlighted on the plots. Dashed line shows genome-wide significance of $P < 5 \times 10^{-8}$. **(A)** Original genome-wide association scan in F508del homozygous patients (minimum $P = 1.4 \times 10^{-9}$ at the initial primary SNP rs483769). **(B)** *P* values for F508del homozygous patients after controlling for the primary SNP showed a minimum at rs7122048 ($P = 2.4 \times 10^{-6}$, regional Bonferroni $P = 0.003$) in the same LD block as the secondary SNP (LD $r^2 = 0.84$ for rs1509661/rs7122048). **(C)** *P* values after controlling for the secondary SNP rs1509661 are much smaller (4-5 orders of magnitude, minimum $P = 6.9 \times 10^{-14}$) than the *P* values from the original scan.

AJRCCM Articles in Press. Published March 15, 2023 as 10.1164/rccm.202209-1653OC

ITGA2 (chr5q11.2)



LAMA2 (chr6q22.3)



SV2B (chr15)

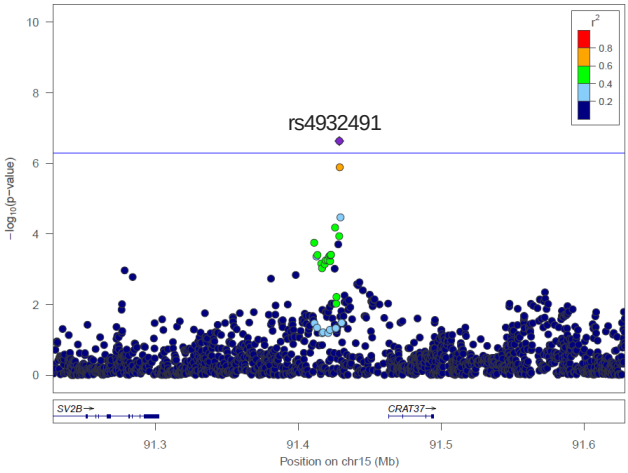
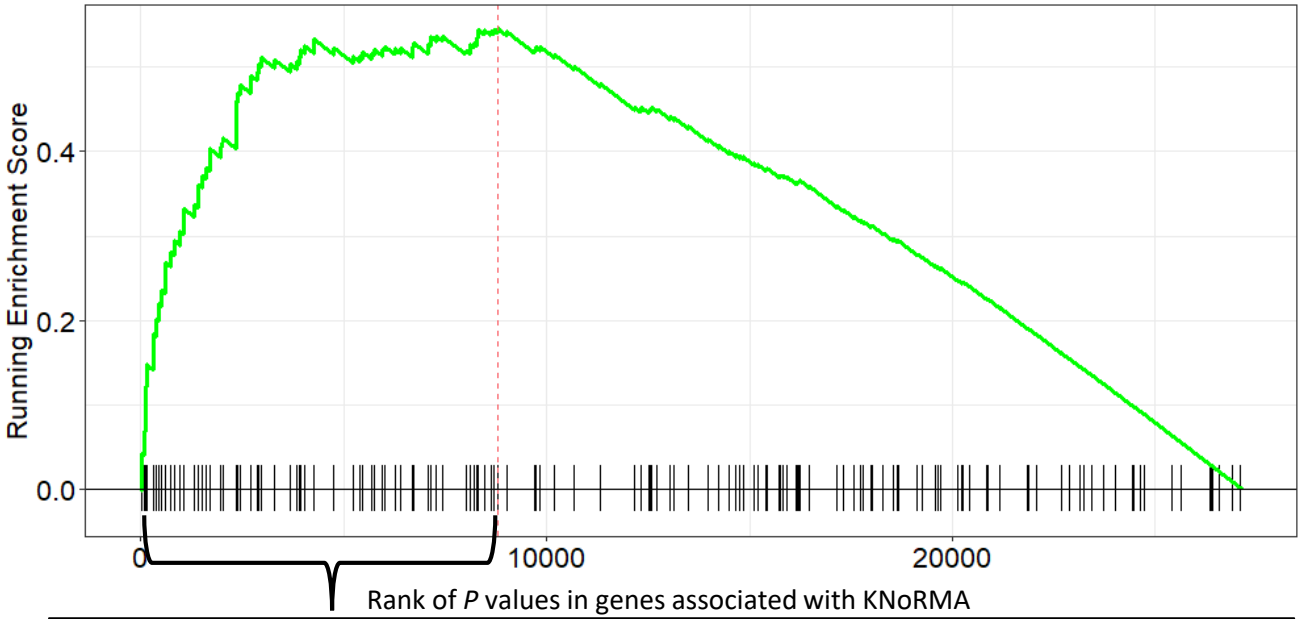


Figure E10. LocusZoom plots for three loci with suggestive associations in the *CFTR* Phe508del/Phe508del homozygotes. Blue line, $P < 5 \times 10\text{-}7$.

AJRCCM Articles in Press. Published March 15, 2023 as 10.1164/rccm.202209-1653OC



RDX, MET, CAPZA2, CARMIL1, MYOC, CDH5, VILL, ADD3, CLASP2, MIR214, KATNB1, TMOD1, SPTBN4, CAPG, KAT2A, CRACD, FRMD7, SPTAN1, CAV3, ARFGEF1, CAPZA3, FLII, F11R, DLC1, CORO1A, LIMA1, CHMP2A, SMAD4, APC2, TWF1, NPM1, FGF13, MAPRE1, TPX2, CCNF, PFN2, BMERB1, MID1, CORO2B, MIR138-1, ATXN7, CAPZB, ARHGAP6, MYADM, ARHGEF18, ESPN, TMSB4X, STMN1, MID1IP1, CAPZA1, GMFG, FHOD3, RBM14, NUBP1, TMEFF2, SSH1, TTBK2, GMFB, CAMSAP2, MKKS, CCP110, PLEKHH2, PPFIA1, SVIL, CLIP3, PAK2, AVIL, SLIT2, INPP5K, SPTBN2, BBS4, ARPIN, SPTBN1, ARHGEF2, KIF25, LMOD1, TACSTD2

Figure E11. Genes that drive core enrichment significant result for this cytoskeleton organization pathway (GO:0051494). This VEGAS2 analysis GSEAS plot includes many genes related to microtubular and cytoskeleton function, which is abnormal in CF epithelial cells (see main text refs 46, 47, and 38).

AJRCCM Articles in Press. Published March 15, 2023 as 10.1164/rccm.202209-1653OC