Published in final edited form as:
Stat Biosci. 2016 ; 2016: 1–17. doi:10.1007/s12561-016-9151-2.

Post-GWAS Prioritization Through Data Integration Provides Novel Insights on Chronic Obstructive Pulmonary Disease

Qiongshi Lu[1], **Chentian Jin**[2], **Jiehuan Sun**[1], **Russell Bowler**[3], **Katerina Kechris**[4], **Naftali Kaminski**[5], and **Hongyu Zhao**[1,6,7,*]

[1]Department of Biostatistics, Yale School of Public Health, New Haven, CT, USA

[2]Yale College, New Haven, CT, USA

[3]National Jewish Health, Department of Medicine, Denver, CO, USA

[4]Department of Biostatistics and Informatics, University of Colorado Denver, Denver, CO, USA

[5]Pulmonary, Critical Care and Sleep Medicine, Yale School of Medicine, New Haven, CT, USA

[6]Program of Computational Biology and Bioinformatics, Yale University, New Haven, CT, USA

[7]VA Cooperative Studies Program Coordinating Center, West Haven, CT, USA

Abstract

Rich collections of genomic and epigenomic annotations, availabilities of large population cohorts for genome-wide association studies (GWAS), and advancements in data integration techniques provide the unprecedented opportunity to accelerate discoveries in complex disease studies through integrative analyses. In this paper, we apply a variety of approaches to integrate GWAS summary statistics of chronic obstructive pulmonary disease (COPD) with functional annotations to illustrate how data integration could help researchers understand complex human diseases. We show that incorporating functional annotations can better prioritize GWAS signals at both the global and the local levels. Signal prioritization on severe COPD GWAS reveals multiple potential risk loci that are linked with pulmonary functions. Enrichment analysis provides novel insights on the pathogenesis of COPD and hints the existence of genetic contributions to muscle dysfuncion and chronic lung inflammation, two symptoms that are often co-morbid with COPD. Our results suggest that rich signals for COPD genetics are still buried under the Bonferroni-corrected genome-wide significance threshold. Many more biological findings are expected to emerge as more samples are recruited for COPD studies.

Introduction

Since its first success more than a decade ago, GWAS has become a popular and powerful approach to study human complex diseases. As of January 15, 2016, more than 15,000 single nucleotide polymorphisms (SNPs) from over 2,000 publications had been documented in the GWAS Catalog (Welter et al. 2014). Despite its great success, GWAS has

[*]Correspondence: Dr. Hongyu, Zhao Department of Biostatistics, Yale School of Public Health, 60 College Street, New Haven, CT, 06511, USA, hongyu.zhao@yale.edu.

several notable limitations. First, complex diseases are often polygenic. Even with thousands or more individuals enrolled in studies, most GWAS are still only able to identify a small fraction of the risk loci with small to moderate effects, and as a result, the loci identified from GWAS only explain a small proportion of disease heritability (Visscher et al. 2012). Second, linkage disequilibrium (LD) allows us to use a few million single nucleotide polymorphisms (SNPs) to tag signals in the whole genome without prior knowledge of where they are, but complex LD structures also hinder our ability to identify functional variants from highly correlated SNPs. Third, nearly 90% of GWAS hits are located in non-protein-coding regions in the human genome (Hindorff et al. 2009). Despite years of efforts through both computational and experimental approaches, interpreting the etiology behind each non-coding risk locus remains challenging (Bernstein et al. 2012; Kellis et al. 2014; Ward and Kellis 2012).

A number of recent studies have shown that incorporating external information (e.g. pleiotropic effect, functional annotation) could accelerate discoveries in many aspects of human genetics research. Joint modeling of multiple diseases have been shown to improve risk stratification for autoimmune and psychiatric disorders (Li et al. 2014; Maier et al. 2015). Genetic correlation studies have revealed novel biological insights for many human complex traits (Bulik-Sullivan et al. 2015; Chung et al. 2014; Pickrell et al. 2015). Integrative analyses of functional annotations and GWAS summary statistics have also been shown to effectively improve GWAS signal identification, prioritization, and interpretation (Finucane et al. 2015; Gusev et al. 2014; Kichaev et al. 2014; Lu et al. 2016a; Lu et al. 2016b; Pickrell 2014).

In this study, we apply a variety of methods to integrate functional annotations of the genome with GWAS summary statistics of COPD, one of the leading causes of morbidity and death in the world (Naghavi et al. 2015). Genetics plays an important role in COPD (Ingebrigtsen et al. 2010). Multiple GWAS have been conducted to identify genetic variants affecting COPD (Cho et al. 2010; Cho et al. 2012; Cho et al. 2014; DeMeo et al. 2009; Pillai et al. 2009; Wilk et al. 2009). To date, five COPD-associated loci have been identified thus far but most heritability for COPD remains unexplained (Zhou et al. 2013). From this perspective, COPD is a perfect phenotype to explore and illustrate the effectiveness of data integration approaches in complex disease research. The results of these integrative analyses could also help researchers generate novel testable hypotheses regarding the etiology of COPD and guide future studies.

Methods

Cohort information

Three sets of summary statistics from two GWAS were used in our analyses. First, we considered the results from the analysis of 5,346 self-described non-Hispanic white individuals participating in COPDGene (Regan et al. 2011), including 2,812 cases and 2,534 controls. These individuals were genotyped using the Illumina HumanOmniExpress array, and genotype imputation was done using 1000 Genomes Phase I v3 European reference panel. The summary statistics of this case-control study are referred to as NHW-casecont in the following sections. Among the cases, 1,390 had grade 3 or 4 disease (severe or very severe) defined by Global Initiative for Chronic Obstructive Lung Disease (GOLD). The summary statistics calculated using this severe sub-cohort as the cases and the same control group are referred to as NHW-severe. Additionally, we also used the summary statistics from a case-control study of 352 African American individuals with severe COPD and 1,749 healthy controls. This dataset is referred to as AA-severe in our following discussion. After filtering out SNPs with low imputation quality (Info < 0.3) and low minor allele frequency (MAF < 0.01), the NHW-casecont, NHW-severe, and AA-severe datasets had 8,119,661, 8,116,544, and 14,214,106 SNPs, respectively. For each SNP, the summary statistic was obtained by applying a logistic regression with adjustment for age, pack-years of smoking, and ancestry-based principal components. Meta-analysis was performed using PLINK (Purcell et al. 2007). Detailed information about all these cohorts was previously reported (Cho et al. 2014).

Functional annotation

GenoCanyon is a statistical framework to predict functional regions in the human genome based on integrative analysis of conservation and epigenomic annotations (Lu et al. 2015). GenoCanyon scores that quantify the functional potential of each nucleotide in the genome have been pre-calculated for the entire hg19 genome. GenoSkyline is a further extension of the GenoCanyon framework through integrating Roadmap Epigenomics data (Kundaje et al. 2015) to quantify tissue-specific functionality in the human genome (Lu et al. 2016a). Currently, GenoSkyline annotations for seven tissue types (i.e. brain, gastrointestinal (GI) tract, lung, heart, blood, muscle, and epithelium) are readily available at http://genocanyon.med.yale.edu/GenoSkyline.

GWAS signal prioritization

We recently developed Genome-Wide Association Prioritizer (GenoWAP), a GWAS signal prioritization approach based on integrative analysis of GWAS summary statistics and functional annotation (Lu et al. 2016b). The GenoWAP algorithm and detailed model assumptions can be found in Lu et al. (2015c). Here we briefly describe the approach for completeness.

For each SNP, define Z and Z_D to be the indicators of general functionality and COPD-specific functionality, respectively. We use p to denote the p-value acquired from standard GWAS analysis. The goal is to use an alternative to the p-value to prioritize all the SNPs. When the GenoCanyon annotation is integrated, non-tissue-specific functionality posterior (NSFP) score, i.e. $P(Z_D = 1|p)$, is used to quantify the importance of each SNP.

NSFP score is calculated using Bayes formula as follows:

$$P(Z_D=1|p)=\frac{f(p|Z_D=1)\times P(Z_D=1)}{f(p|Z_D=1)\times P(Z_D=1)+f(p|Z_D=0)\times P(Z_D=0)} \quad (1)$$

where $P(Z_D = 1)$ can be further denoted as:

$$P(Z_D=1)=P(Z_D=1|Z=1)\times P(Z=1) \quad (2)$$

In this formulation, $P(Z = 1)$ is defined as the average GenoCanyon score of the surrounding 10,000 base pairs for each SNP. First, we partition all the SNPs into two subgroups based on a mean GenoCanyon score cutoff of 0.1. In this way, $f(p|Z_D = 0)$ can be directly estimated by applying density estimation techniques on the SNP subgroup with low annotation scores. More specifically, a histogram method is used for density estimation and the optimal number of bins is chosen through cross-validation. Second, we assume that, regardless of local LD structure, SNPs that are not relevant to the phenotype will have similar p-value behavior to the SNPs that are not annotated to be functional. More formally, we can describe this relationship as follows:

$$f(p|Z_D=0)=f(p|Z=0) \quad (3)$$

The above equation essentially assumes that LD, the major driving force of nonuniformity in $(p|Z_D = 0)$ and $(p|Z = 0)$, has the same impact on both $f(p|Z_D = 0)$ and $f(p|Z = 0)$. Indeed, it has been previously shown that the LD pattern in the nonfunctional genome is not significantly different from that in the whole genome (Lu et al. 2016b).

Finally, we estimate all the remaining terms in equations 1 and 2, including $f(p|Z_D = 1)$ and $P(Z_D = 1|Z =1)$, using the EM algorithm. In the first step of the estimation procedure, we acquired the subset of SNPs located in functional regions. The p-value distribution of these SNPs is the following mixture.

$$\begin{aligned} f(p|Z=1) &= P(Z_D=1|Z=1)\times f(p|Z_D=1,Z=1)+P(Z_D=0|Z=1)\times f(p|Z_D=0,Z=1)\\ &= P(Z_D=1|Z=1)\times f(p|Z_D=1)+P(Z_D=0|Z=1)\times f(p|Z=0) \end{aligned} \quad (4)$$

The last equality is the consequence of assumption (3) and the definitions of Z and Z_D. Density $f(p|Z = 0)$ has been estimated in earlier steps. Similar to Chung et al. (2014), we assume a beta distribution of the p-values of functional SNPs, i.e. $f(p|Z_D = 1)$ as a reasonable approximation under some assumptions of SNP effect size.

$$(p|Z_D=1)\sim Beta(\alpha,1), \quad 0<\alpha<1 \quad (5)$$

The EM algorithm is then applied to the SNP subset located in the functional genome, i.e. $Z = 1$. The beta assumption guarantees a closed-form expression in each iteration and all the remaining parameters can be subsequently estimated.

LD score regression

LD score regression (Finucane et al. 2015) was used to estimate the signal enrichment in SNP categories based on functional annotation. First, annotation-stratified LD scores were computed using GenoSkyline annotations (Lu et al. 2016a), 1000 Genomes data of European ancestry (Abecasis et al. 2012), and a 1-centiMorgan window. Then, annotation-stratified LD scores of seven tissue types were jointly analyzed using LD score regression. Enrichment was calculated based on the ratio of explained heritability and the proportion of SNPs in each annotation category.

Software availability

Implemented GenoWAP software is accessible at our server (http://genocanyon.med.yale.edu/GenoWAP). A detailed user manual for GWAS signal prioritization is also provided. Required files for enrichment analysis using LD score regression and sample codes can be accessed at http://genocanyon.med.yale.edu/GenoSkyline.

Results

Prioritizing signals at GWAS loci for COPD and pulmonary function

NSFP scores for all 8,119,661 SNPs in the NHW-casecont dataset were calculated using GenoWAP software (Supplementary Figure 1). P-value distributions for SNPs in the functional and non-functional regions (i.e. $f(p|Z=1)$ and $f(p|Z=0)$; **Methods**) are shown in Figure 1. Signal enrichment in the functional genome was highly significant (p-value = 2.69×10^{-42}, one-sided Kolmogorov-Smirnov test).

Five risk loci have been identified and replicated in previously published GWAS for COPD. In order to compare the signal prioritization performance of different metrics, ranks of the lowest p-value and the largest NSFP score at each locus are compared (Table 1). Four out of five loci have a tied or improved rank using NSFP score. The locus upstream of *HHIP* on chromosome 4 is the only locus with a decreased rank. This locus will be discussed in detail later.

Two large-scale GWAS (Hancock et al. 2010; Repapi et al. 2010) identified 11 loci associated with pulmonary function (i.e. FEV_1 and FVC). Later, a joint GWAS meta-analysis (Artigas et al. 2011) further increased the number of risk loci to 27. We compared the p-value-based and NSFP-based ranks of these 27 loci (Table 2). 22 out of 27 loci showed improved ranks (p-value = 7.57×10^{-4}, one-sided binomial test). Only four loci had decreased ranks, among which two loci only showed up in the joint meta-analysis. The locus on chromosome 2q36 had a drastic decline in its rank under NSFP score. Interestingly, it was also the only locus that was not successfully replicated in the meta-analysis (Artigas et al. 2011).

Prioritizing signals at *HHIP* locus

The intergenic region upstream of *HHIP* on chromosome 4 has been repeatedly identified in multiple lung function GWAS (Artigas et al. 2011; Hancock et al. 2010; Repapi et al. 2010) and several COPD studies (Cho et al. 2012; Cho et al. 2014). Moreover, the signal pattern at this locus is highly consistent across different studies. A signal plateau that spans approximately 300Kb (145.25 – 145.55Mb; hg19) could be observed in all these studies. Even more interestingly, the 100Kb region (145.45 – 145.55Mb) near the transcription start site of *HHIP* always showed the strongest signal in this LD block. However, the same 100Kb signal peak could not be observed in the NHW-casecont dataset (Figure 2). Instead, several SNPs near the middle of this LD block (e.g. rs1032295; chr4: 145,434,584; p-value = 4.66×10^{-7}) showed substantially stronger signals than SNPs in the 100Kb region.

Interestingly, the 100Kb signal peak re-appeared after signal prioritization. NSFP scores clearly suggested that the region between 145.45 and 145.55Mb is most likely the signal source at this locus. Furthermore, rs13141641, the SNP with the largest NSFP score, showed the lowest p-value at this locus in two previously published COPD studies (Cho et al. 2012; Cho et al. 2014). In this dataset, however, its raw p-value was unimpressive (p-value = 1.02×10^{-5}), showing substantial gaining of power when integrating external information. This also explains why this locus had a lower rank under NSFP score than p-value as stated above. In table 1, the p-value-based rank of SNP rs1032295 was compared with the NSFP-based rank of SNP rs13141641. Since the raw signal at rs13141641 was only moderate, the NSFP score remained moderate after annotation integration. Locally, however, it was already sufficient to remove noises due to LD and reveal the truly functional element from correlated neighboring SNPs.

Despite its strong association with pulmonary function and COPD, the functional mechanism at this locus was not clear until chromosome conformation capture (3C) experiment suggested its functional impact as a *HHIP* enhancer (Zhou et al. 2012). Among the four distant 3C fragments being tested, one fragment (145,481,550 – 145,488,550; hg19) showed strong physical interaction with the *HHIP* promoter. This fragment is completely contained in the 100Kb signal peak favored by our signal prioritization approach. Notably, another peak at 145.25 – 145.29Mb also survived signal prioritization (Figure 2), yet none of the fragments in the 3C study covered this region (Zhou et al. 2012). Its functional impact remains to be investigated in the future.

Identifying genetic loci associated with severe COPD using NSFP score

The COPDGene cohort has been used in several large-scale GWAS meta-analyses, but signals for severe COPD cases are relatively less explored. The only GWAS that focused on severe COPD replicated signals at four previously established COPD risk loci (*CHRNA3*, *FAM13A*, *HHIP*, and *RIN3*) and identified two additional loci (*MMP3* and *TGFB2*). In this section, we explore highly ranked loci based on NSFP scores in the NHW-severe cohort and seek signal replication using the AA-severe cohort.

GenoWAP was applied to the NHW-Severe summary statistics (Supplementary Figure 2). Signals were significantly enriched in the functional genome (p-value = 4.18×10^{-68}, one-sided Kolmogorov-Smirnov test). 233 SNPs from 14 loci had NSFP scores greater than 0.7. This cutoff is comparable to the p-value cutoff of 1×10^{-6} in this dataset, which yields 239 SNPs. However, the 239 SNPs based on p-value could only cover 4 loci, all of which were included in the 14 loci based on NSFP score. Remarkably, compared with the estimated effect sizes in the NHW-casecont dataset, all 233 SNPs have stronger estimated effect sizes in the GWAS of severe COPD (Supplementary Table 1). In order to replicate signals in an independent cohort, we extracted p-values for these 233 SNPs from the AA-Severe dataset. 39 were significant under cutoff 0.05. Signals were successfully replicated at four previously reported COPD risk loci, and one novel locus on chromosome 20q11.21 (Table 3). Three loci (*CHRNA3*, *HHIP*, and *RIN3*) achieved genome-wide significance after random-effect meta-analysis. One limitation of this two-stage trans-ethnic analysis strategy is its inability to identify population-specific COPD risk loci. Therefore, the unreplicated loci remain to be further investigated using independent cohorts with European ancestry.

Both NHW-Severe and AA-Severe cohorts have been used in the GWAS meta-analysis of severe COPD published in 2014 (Cho et al. 2014). Therefore, it is not surprising that no novel locus achieved genome-wide significance in our analysis. However, a few loci with large NSFP scores are related with pulmonary function, and may be replicated in future studies using a larger sample size (Supplementary Table 1). First, signals at the *POFUT1-PLAGL2* locus on chromosome 20 were replicated in the African American cohort (Table 3 and Figure 3). *POFUT1* encodes an O-fucosyltransferase essential for NOTCH signaling, which plays a crucial role in development and homeostasis of the lung (Xu et al. 2012). The pathogenic role of *PLAGL2* in pulmonary emphysema, a major component of severe COPD, has also been established (Yang et al. 2009). Despite not replicated, a few other loci also provided some interesting insights. SNP rs7664805 at the *GSTCD-NPNT* locus had a very high NSFP score (0.9352) in the NHW-severe dataset. This locus had been repeatedly identified as an associated locus of FEV_1 (Artigas et al. 2011; Hancock et al. 2010; Repapi et al. 2010), a commonly used indicator of pulmonary function (Table 2). Gene *SFTPC* on chromosome 4 encodes surfactant protein C, a protein that is essential for pulmonary function and homeostasis. Variants in *SFTPC* have been linked with interstitial lung disease, lung function, as well as obstructive lung disease (Baekvad-Hansen et al. 2010). Finally, *EGLN2* has been previously identified as a COPD-associated locus (Cho et al. 2012). One of its important paralogs, *EGLN3*, also showed up in our analysis.

Finally, for completeness of our analysis, we applied GenoWAP to the fixed-effect meta-analysis results combining NHW-Severe and AA-Severe cohorts. Due to the absence of replication cohort, we chose a more stringent NSFP cutoff of 0.9, which gave 106 SNPs. This threshold is comparable to the p-value cutoff of 1×10^{-7} in the meta-analysis, which would give 108 SNPs. Seven loci were identified using this NSFP cutoff (Supplementary Table 2), from which five loci (*HHIP*, *FAM13A*, *MMP12*, *RIN3*, *CHRNA3*) have been previously reported in the severe COPD GWAS (Cho et al. 2014). The *POFUT1-PLAGL2* locus has been identified and replicated in our analysis above. The other locus, TRIM2 on chromosome 4q31.3, is a novel finding. Its association with COPD or lung function has never been identified before. The validity of these signals requires further replication in independent and larger cohorts. If validated, these results could help researchers generate testable hypotheses regarding the etiology of COPD.

Partitioning heritability by annotation categories

Signal enrichment in annotated categories can greatly help researchers understand complex diseases. Recently, Finucane et al. (2015) proposed to use LD score regression to partition heritability by functional annotations and identified some surprising and interesting results, e.g. strong enrichment of BMI associations in central nervous system (Finucane et al. 2015). We applied LD score regression to the NHW-casecont summary statistics and estimated the signal enrichment in seven tissue-specific functional categories (**Methods**). Interestingly, muscle was significantly enriched of COPD signals and it had the largest fold enrichment (p-value = 0.011, enrichment = 10.004; Figure 4 and Supplementary Table 3). Exercise intolerance is one of the major problems in COPD patients (Casaburi 2001). It has long been recognized that alterations in skeletal muscle function and structure, independent of pulmonary function, significantly limits exercise capacity (Wüst and Degens 2007). It has also been established that many systemic and local factors exert strong effects on skeletal muscle dysfunction in COPD patients (Kim et al. 2008). Our enrichment result hints the existence of genetic contribution.

Besides muscle, blood-specific functional category is also enriched of COPD associations. In fact, despite the lower fold change (enrichment = 7.440), the enrichment was even more significant (p-value = 0.009). It is widely recognized that COPD is an inflammatory diseases of the airways and immunity plays a central role in COPD (Bhat et al. 2015; Rovina et al. 2013). Our results show that genetic variants located in immune-related regions may be involved in the etiology of COPD. Notably, the enrichment in lung was not significant (p-value = 0.147). However, the estimated fold enrichment was high (enrichment = 8.533; Figure 4). The large estimate of standard error is partly due to the low proportion SNPs covered by lung annotation (Supplementary Table 3). Whether enrichment in lung exists or not needs to be further investigated using summary statistics of larger cohorts. We also estimated signal enrichment in the 53 baseline annotations of LD score regression (Supplementary Table 4). Significantly enriched categories include super enhancers, histone mark H3K4me1, and histone mark H3K9ac. These results suggest a crucial role of non-coding regulatory regions in COPD etiology. On the contrary, signals are significantly depleted in the regressed genome.

Discussion

In this paper, we have applied various methods to integrate COPD GWAS summary statistics with genomic functional annotations, and illustrated how these integrative approaches could benefit complex disease research. Globally, the ranks of well-established risk loci based on NSFP score are substantially higher than the ranks based on p-values. Locally, NSFP score identified the *HHIP* enhancer region within a large LD block on chromosome 4. These results demonstrate that functional annotations could effectively reduce the noise due to chance and LD structure. Integrating GWAS summary statistics with annotation data could better prioritize signals at both global and local levels.

We also used annotation data to prioritize the GWAS signals of severe COPD. It is not surprising that no novel risk locus reached genome-wide significance because these data have been used in previously published GWAS meta-analysis. However, several top loci based on NSFP score have known functions directly or indirectly related to the etiology of COPD. GWAS loci for complex diseases usually have small to moderate effect sizes. Different functional variants and distinct LD structure across populations also make it challenging to replicate signals in trans-ethnic analysis. Therefore, whether these signals are true positives still remains to be investigated using large and homogeneous cohorts of severe COPD.

Enrichment of associations in annotated categories has brought novel insights on many human complex diseases and traits. Interesting results were also seen for COPD. Significant enrichment in muscle and blood is in agreement with the involvement of skeletal muscle dysfunction and lung inflammation in COPD, and these results hint a substantial genetic contribution to these symptoms. However, LD score regression works better with large sample size and strong signals in the GWAS dataset. These results as well as the enrichment in other tissue types could be further validated using a larger cohort in the future.

In summary, integrative analysis of COPD GWAS summary statistics and functional annotations revealed interesting signals in the current GOPDGene cohort. Many more novel and insightful results should be expected as the sample size rapidly grows. Integrative approaches that bring in external information from annotation data have been shown to greatly benefit GWAS as well as many other aspects of human genetics research. In the era of “big data”, with more and more high-quality annotation data becoming available, biologically motivated and statistically sound approaches will play a central role in omics data integration and continue to accelerate discoveries in complex disease studies.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank all members of the Statistical and Applied Mathematical Sciences Institute (SAMSI) Data Integration - COPD Working Group as part of the SAMSI Beyond Bioinformatics Program. We also thank Dr. Michael Cho for sharing COPDgene GWAS summary statistics. We are grateful for the support of Drs. Sujit Ghosh and Snehalata Huzurbazar at SAMSI. This work was supported in part by the National Institutes of Health grant R01 GM59507 and the VA Cooperative Studies Program of the Department of Veterans Affairs, Office of Research and Development. Funding for the COPDGene Study is provided by the National Heart, Lung, And Blood Institute Award Number R01HL089897 and Award Number R01HL089856. Further support provided by the COPD Foundation through contributions made to an Industry Advisory Board comprising AstraZeneca, Boehringer Ingelheim, Novartis Pfizer, Siemens, and Sunovion.

References

Abecasis GR, et al. An integrated map of genetic variation from 1,092 human genomes. Nature. 2012; 491:56–65. DOI: 10.1038/nature11632 [PubMed: 23128226]

Artigas MS, et al. Genome-wide association and large-scale follow up identifies 16 new loci influencing lung function. Nature genetics. 2011; 43:1082–1090. [PubMed: 21946350]

Baekvad-Hansen M, Nordestgaard BG, Tybjaerg-Hansen A, Dahl M. Two novel mutations in surfactant protein-C, lung function and obstructive lung disease. Respiratory medicine. 2010; 104:418–425. DOI: 10.1016/j.rmed.2009.10.012 [PubMed: 19910179]

Bernstein BE, Birney E, Dunham I, Green ED, Gunter C, Snyder M. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012; 489:57–74. DOI: 10.1038/nature11247 [PubMed: 22955616]

Bhat TA, Panzica L, Kalathil SG, Thanavala Y. Immune Dysfunction in Patients with Chronic Obstructive Pulmonary Disease. Annals of the American Thoracic Society. 2015; 12:S169–S175. DOI: 10.1513/AnnalsATS.201503-126AW [PubMed: 26595735]

Bulik-Sullivan B, et al. An atlas of genetic correlations across human diseases and traits. Nature genetics. 2015; 47:1236–1241. DOI: 10.1038/ng.3406 [PubMed: 26414676]

Casaburi R. Skeletal muscle dysfunction in chronic obstructive pulmonary disease. Medicine and science in sports and exercise. 2001; 33:S662–670. [PubMed: 11462075]

Cho MH, et al. Variants in FAM13A are associated with chronic obstructive pulmonary disease. Nature genetics. 2010; 42:200–202. DOI: 10.1038/ng.535 [PubMed: 20173748]

Cho MH, et al. A genome-wide association study of COPD identifies a susceptibility locus on chromosome 19q13. Human molecular genetics. 2012; 21:947–957. [PubMed: 22080838]

Cho MH, et al. Risk loci for chronic obstructive pulmonary disease: a genome-wide association study and meta-analysis. The lancet Respiratory medicine. 2014; 2:214–225. DOI: 10.1016/s2213-2600(14)70002-5 [PubMed: 24621683]

Chung D, Yang C, Li C, Gelernter J, Zhao H. GPA: A Statistical Approach to Prioritizing GWAS Results by Integrating Pleiotropy and Annotation. PLoS genetics. 2014; 10:e1004787.doi: 10.1371/journal.pgen.1004787 [PubMed: 25393678]

DeMeo DL, et al. Integration of genomic and genetic approaches implicates IREB2 as a COPD susceptibility gene. The American Journal of Human Genetics. 2009; 85:493–502. [PubMed: 19800047]

Finucane HK, et al. Partitioning heritability by functional annotation using genome-wide association summary statistics. Nature genetics. 2015

Gusev A, et al. Partitioning heritability of regulatory and cell-type-specific variants across 11 common diseases. The American Journal of Human Genetics. 2014; 95:535–552. [PubMed: 25439723]

Hancock DB, et al. Meta-analyses of genome-wide association studies identify multiple loci associated with pulmonary function. Nature genetics. 2010; 42:45–52. [PubMed: 20010835]

Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, Manolio TA. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proceedings of the National Academy of Sciences of the United States of America. 2009; 106:9362–9367. DOI: 10.1073/pnas.0903103106 [PubMed: 19474294]

Ingebrigtsen T, et al. Genetic influences on chronic obstructive pulmonary disease–A twin study. Respiratory medicine. 2010; 104:1890–1895. [PubMed: 20541380]

Kellis M, et al. Defining functional DNA elements in the human genome. Proceedings of the National Academy of Sciences of the United States of America. 2014; doi: 10.1073/pnas.1318948111

Kichaev G, et al. Integrating Functional Data to Prioritize Causal Variants in Statistical Fine-Mapping Studies. PLoS genetics. 2014; 10:e1004722.doi: 10.1371/journal.pgen.1004722 [PubMed: 25357204]

Kim HC, Mofarrahi M, Hussain SN. Skeletal muscle dysfunction in patients with chronic obstructive pulmonary disease. International journal of chronic obstructive pulmonary disease. 2008; 3:637. [PubMed: 19281080]

Kundaje A, et al. Integrative analysis of 111 reference human epigenomes. Nature. 2015; 518:317–330. [PubMed: 25693563]

Li C, Yang C, Gelernter J, Zhao H. Improving genetic risk prediction by leveraging pleiotropy. Human genetics. 2014; 133:639–650. [PubMed: 24337655]

Lu Q, Hu Y, Sun J, Cheng Y, Cheung K-H, Zhao H. A Statistical Framework to Predict Functional Non-Coding Regions in the Human Genome Through Integrated Analysis of Annotation Data. Sci Rep. 2015; 5doi: 10.1038/srep10576

Lu Q, Powles RL, Wang Q, He BJ, Zhao H. Integrative Tissue-Specific Functional Annotations in the Human Genome Provide Novel Insights on Many Complex Traits and Improve Signal Prioritization in Genome Wide Association Studies. PLoS genetics. 2016a; 12:e1005947.doi: 10.1371/journal.pgen.1005947 [PubMed: 27058395]

Lu Q, Yao X, Hu Y, Zhao H. GenoWAP: GWAS signal prioritization through integrated analysis of genomic functional annotation. Bioinformatics. 2016b; 32:542–548. DOI: 10.1093/bioinformatics/btv610 [PubMed: 26504140]

Maier R, et al. Joint analysis of psychiatric disorders increases accuracy of risk prediction for schizophrenia, bipolar disorder, and major depressive disorder. The American Journal of Human Genetics. 2015; 96:283–294. [PubMed: 25640677]

Naghavi M, et al. Global, regional, and national age-sex specific all-cause and cause-specific mortality for 240 causes of death, 1990–2013: a systematic analysis for the Global Burden of Disease Study 2013. Lancet. 2015; 385:117–171. [PubMed: 25530442]

Pickrell J, Berisa T, Segurel L, Tung JY, Hinds D. Detection and interpretation of shared genetic influences on 40 human traits. 2015 bioRxiv:019885.

Pickrell JK. Joint analysis of functional genomic data and genome-wide association studies of 18 human traits. The American Journal of Human Genetics. 2014; 94:559–573. [PubMed: 24702953]

Pillai SG, et al. A genome-wide association study in chronic obstructive pulmonary disease (COPD): identification of two major susceptibility loci. PLoS genetics. 2009; 5:e1000421. [PubMed: 19300482]

Pruim RJ, et al. LocusZoom: regional visualization of genome-wide association scan results. Bioinformatics. 2010; 26:2336–2337. [PubMed: 20634204]

Purcell S, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. The American Journal of Human Genetics. 2007; 81:559–575. [PubMed: 17701901]

Regan EA, et al. Genetic epidemiology of COPD (COPDGene) study design. COPD: Journal of Chronic Obstructive Pulmonary Disease. 2011; 7:32–43.

Repapi E, et al. Genome-wide association study identifies five loci associated with lung function. Nature genetics. 2010; 42:36–44. DOI: 10.1038/ng.501 [PubMed: 20010834]

Rovina N, Koutsoukou A, Koulouris NG. Inflammation and immune response in COPD: where do we stand? Mediators of inflammation. 2013

Visscher PM, Brown MA, McCarthy MI, Yang J. Five years of GWAS discovery. American journal of human genetics. 2012; 90:7–24. DOI: 10.1016/j.ajhg.2011.11.029 [PubMed: 22243964]

Ward LD, Kellis M. Interpreting noncoding genetic variation in complex traits and human disease. Nature biotechnology. 2012; 30:1095–1106. DOI: 10.1038/nbt.2422

Welter D, et al. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. Nucleic Acids Res. 2014; 42:D1001–1006. DOI: 10.1093/nar/gkt1229 [PubMed: 24316577]

Wilk JB, et al. A genome-wide association study of pulmonary function measures in the Framingham Heart Study. PLoS genetics. 2009; 5:e1000429. [PubMed: 19300500]

Wüst RC, Degens H. Factors contributing to muscle wasting and dysfunction in COPD patients. International journal of chronic obstructive pulmonary disease. 2007; 2:289. [PubMed: 18229567]

Xu, K.; Moghal, N.; Egan, SE. Notch Signaling in Embryology and Cancer. Springer; 2012. Notch signaling in lung development and disease; p. 89-98.

Yang Y-S, Yang M-CW, Guo Y, Williams OW, Weissler JC. PLAGL2 expression-induced lung epithelium damages at bronchiolar alveolar duct junction in emphysema: bNip3-and SP-C-associated cell death/injury activity. American Journal of Physiology-Lung Cellular and Molecular Physiology. 2009; 297:L455–L466. [PubMed: 19574421]

Zhou JJ, Cho MH, Castaldi PJ, Hersh CP, Silverman EK, Laird NM. Heritability of chronic obstructive pulmonary disease and related phenotypes in smokers. American journal of respiratory and critical care medicine. 2013; 188:941–947. [PubMed: 23972146]

Zhou X, et al. Identification of a chronic obstructive pulmonary disease genetic determinant that regulates. HHIP Human molecular genetics. 2012; 21:1325–1335. DOI: 10.1093/hmg/ddr569 [PubMed: 22140090]

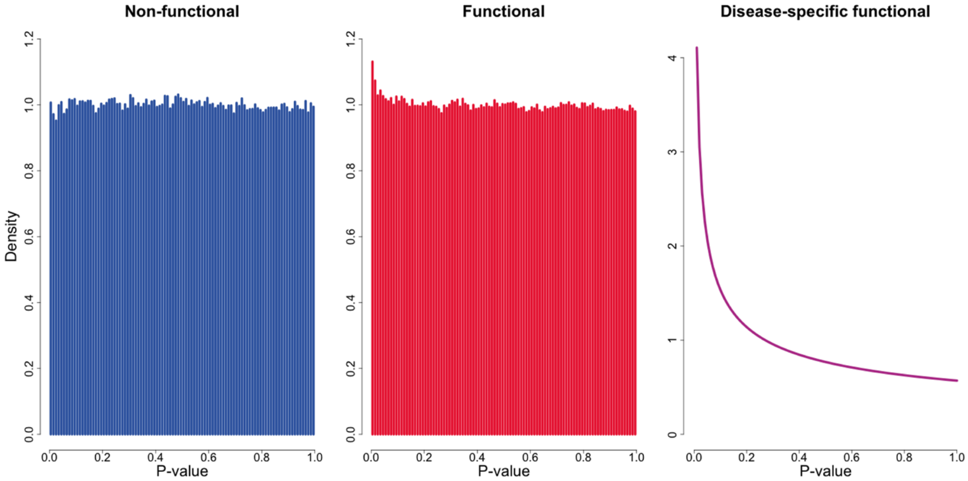

Figure 1. P-value distributions for different SNP categories in the NHW-casecont dataset

From left to right, the three panels show the p-value distributions for non-functional (Z=0), functional (Z=1), and disease-specific functional (Z_D=1) categories, respectively.

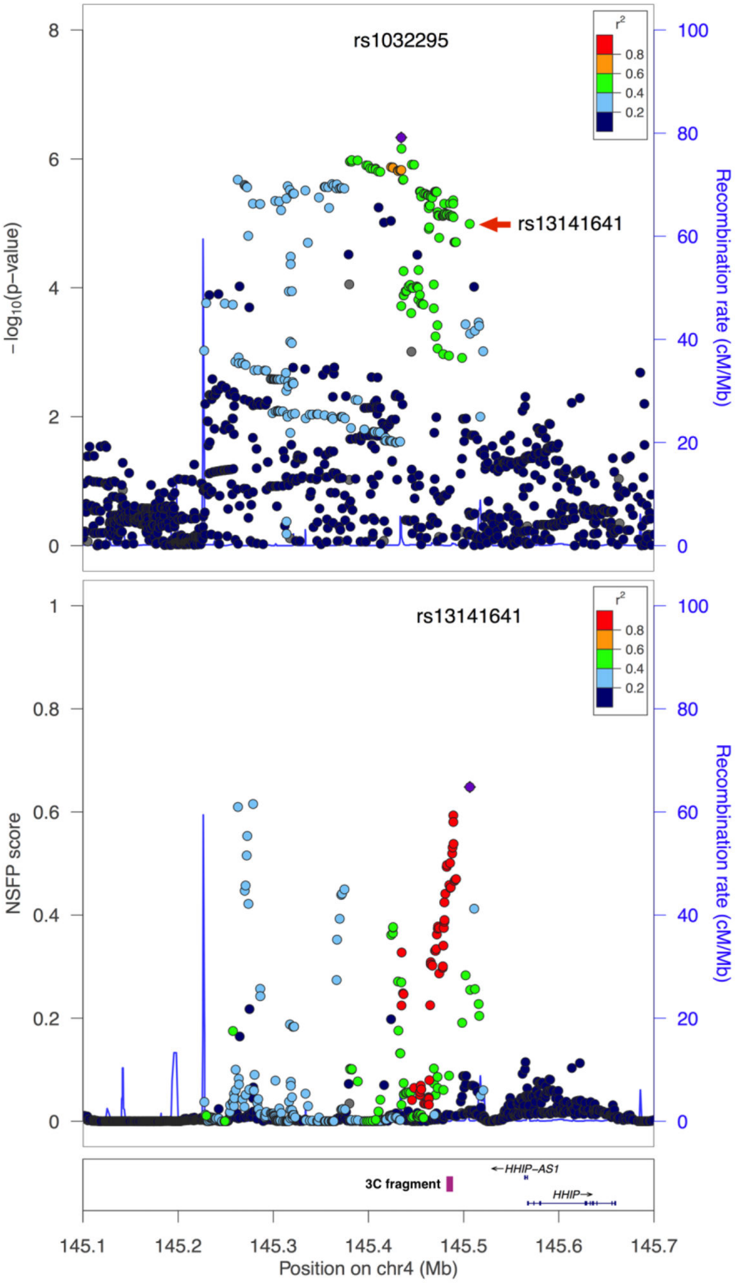

Figure 2. Signals at the risk locus upstream of *HHIP*

The upper panel shows the p-values in the NHW-casecont dataset. The lower panel shows the NSFP scores. The 3C fragment with strong physical interactions with the *HHIP* promoter is highlighted in a purple box. Locus plots were made using LocusZoom (Pruim et al. 2010).

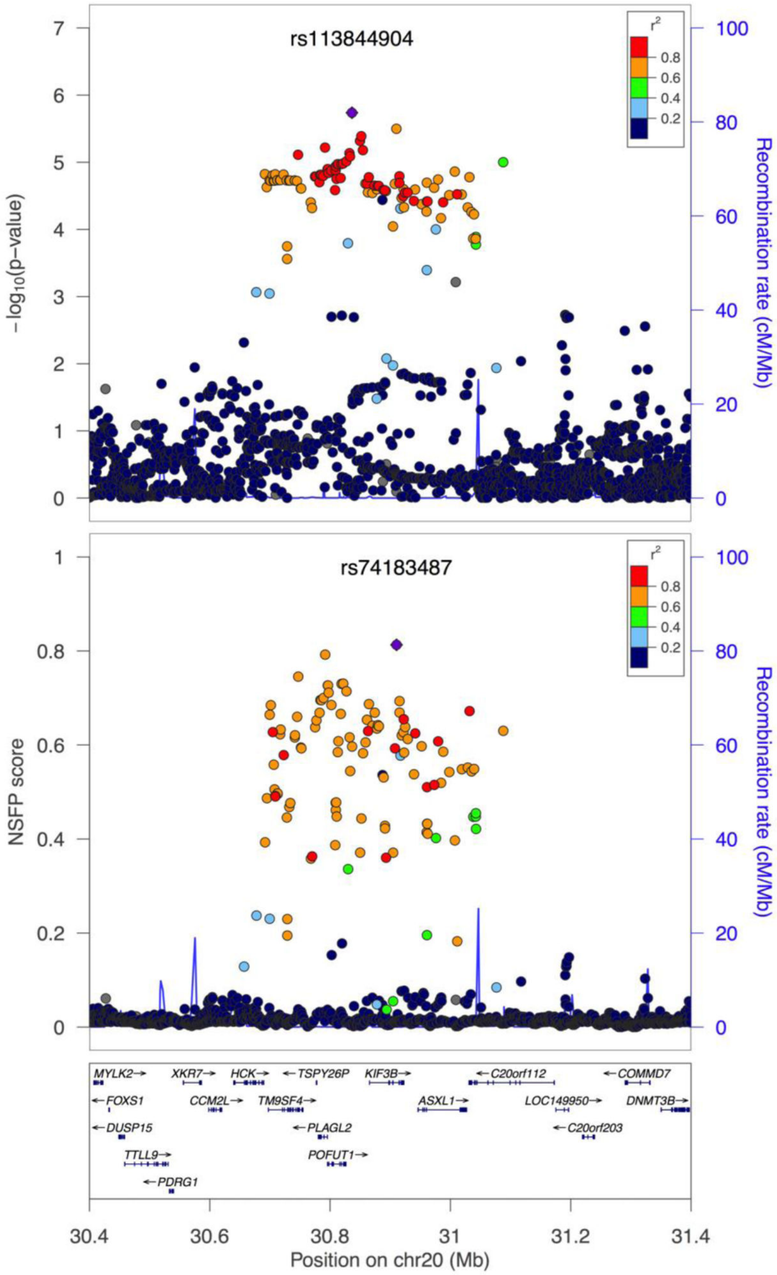

Figure 3. Signals at a potential risk locus on chromosome 20

The upper panel shows the p-values in the NHW-severe dataset. The lower panel shows the NSFP scores. Locus plots were made using LocusZoom (Pruim et al. 2010).

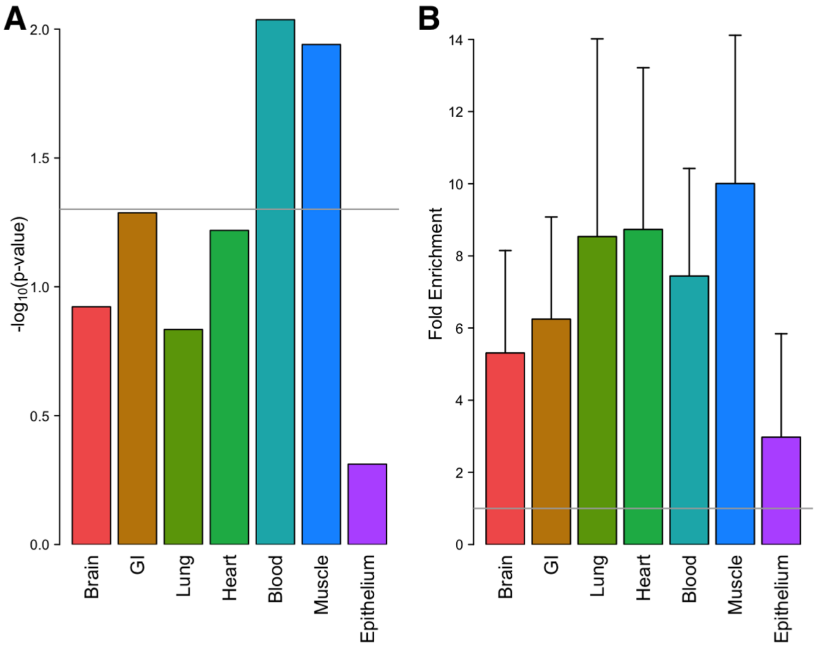

Figure 4. Enrichment estimates for GenoSkyline annotations

(A) Log-transformed p-values for seven tissue types. The grey line indicates the p-value cutoff of 0.05. (B) Fold enrichments and standard errors. The grey line indicates enrichment=1.

Table 1

Previously reported COPD-associated loci.

			Top signals based on p-value			Top signals based on NSFP score		
Chr	**Gene[1]**	**Ref**	**Leading SNP**	**P-value**	**Rank[2]**	**Leading SNP**	**NSFP**	**Rank[3]**
4q22	*FAM13A*	(Cho et al. 2010; Cho et al. 2012; Cho et al. 2014)	rs4416442	1.85E-08	2	rs6837671	0.648	2
4q31	*HHIP*	(Cho et al. 2010; Cho et al. 2012; Cho et al. 2014; Pillai et al. 2009)	rs1032295	4.66E-07	20	rs13141641	0.802	135
14q32	*RIN3*	(Cho et al. 2014)	rs72699866	1.65E-06	79	rs72699866	0.942	15
15q25	*CHRNA3*	(Cho et al. 2010; Cho et al. 2012; Cho et al. 2014; Pillai et al. 2009)	rs8040868	1.30E-08	1	rs8040868	0.909	1
19q13	*EGLN2*	(Cho et al. 2012)	rs193122850	7.85E-05	1237	rs193122850	0.433	527

[1] Only one gene at each locus is listed.

[2] rank of the SNP with the smallest p-value at each locus

[3] rank of the SNP with the largest NSFP score at each locus

Table 2

Previously reported risk loci for pulmonary function.

			Top signals based on p-value			Top signals based on NSFP score		
Chr	**Gene[1]**	**Ref**	**Leading SNP**	**P-value**	**Rank[2]**	**Leading SNP**	**NSFP**	**Rank[3]**
2q35	*TNS1*	(Artigas et al. 2011; Repapi et al. 2010)	rs75596991	1.36E-02	118351	rs75596991	0.080	30358
4q24	*GSTCD*	(Artigas et al. 2011; Hancock et al. 2010; Repapi et al. 2010)	rs7664805	3.14E-05	739	rs7664805	0.392	626
4q31	*HHIP*	(Artigas et al. 2011; Hancock et al. 2010; Repapi et al. 2010)	rs1032295	4.66E-07	20	rs13141641	0.648	135
5q33	*HTR4*	(Artigas et al. 2011; Hancock et al. 2010; Repapi et al. 2010)	rs4597955	4.18E-05	831	rs7730971	0.462	454
6p21	*AGER*	(Artigas et al. 2011; Hancock et al. 2010; Repapi et al. 2010)	rs2070600	1.90E-05	592	rs2070600	0.586	241
15q23	*THSD4*	(Artigas et al. 2011; Repapi et al. 2010)	rs12899576	6.99E-04	7393	rs12899576	0.229	2216
6q24	*GPR126*	(Artigas et al. 2011; Hancock et al. 2010)	rs17280293	1.04E-03	10501	rs17280293	0.202	3019
5q33	*ADAM19*	(Artigas et al. 2011; Hancock et al. 2010)	rs79101970	1.88E-04	2279	rs190229691	0.295	1180
4q22	*FAM13A*	(Artigas et al. 2011; Hancock et al. 2010)	rs4416442	1.85E-08	2	rs6837671	0.909	2
9q22	*PTCH1*	(Artigas et al. 2011; Hancock et al. 2010)	rs2405373	2.66E-03	24951	rs28536742	0.129	8946
2q36	*PID1*	(Hancock et al. 2010)	rs16825267	2.56E-04	2957	rs4972871	0.071	40480
1p36	*MFAP2*	(Artigas et al. 2011)	rs34961969	3.83E-03	35369	rs115686702	0.105	14626
1q41	*TGFB2*	(Artigas et al. 2011)	rs3009947	2.13E-03	20316	rs4846478	0.139	7562
2q37	*HDAC4*	(Artigas et al. 2011)	rs12692237	4.24E-04	4670	rs12692237	0.234	2108
3p24	*RARB*	(Artigas et al. 2011)	rs186061478	1.83E-03	17604	rs111515348	0.103	15705
3q26	*MECOM*	(Artigas et al. 2011)	rs111575942	1.15E-03	11350	rs960354	0.126	9422
5q15	*SPATA9*	(Artigas et al. 2011)	rs187403530	1.17E-03	11519	rs187403530	0.194	3332
6p22	*ZKSCAN3*	(Artigas et al. 2011)	rs200991	1.35E-02	117511	rs200956	0.070	41746
6p21	*NCR3*	(Artigas et al. 2011)	rs2844456	4.42E-04	4813	rs2844456	0.266	1540
6q21	*ARMC2*	(Artigas et al. 2011)	rs2848598	9.22E-03	80934	rs2806356	0.082	28615
10p23	*CDC123*	(Artigas et al. 2011)	rs75346809	3.57E-03	33172	rs11815176	0.096	18269
10q22	*C10orf11*	(Artigas et al. 2011)	rs78051646	5.52E-04	5947	rs989301	0.143	7074
12q13	*LRP1*	(Artigas et al. 2011)	rs185801064	4.04E-04	4489	rs138854007	0.251	1765
12q22	*CCDC38*	(Artigas et al. 2011)	rs146931557	5.14E-03	46765	rs9971896	0.088	23355
16q13	*MMP15*	(Artigas et al. 2011)	rs75625208	1.67E-03	16253	rs78390789	0.149	6406
16q23	*CFDP1*	(Artigas et al. 2011)	rs11645953	5.73E-04	6174	rs11645953	0.149	6357
21q22	*KCNE2*	(Artigas et al. 2011)	rs184449084	1.66E-03	16175	rs184449084	0.164	4976

[1] Only one gene at each locus is listed.

[2] rank of the SNP with the smallest p-value at each locus

[3] rank of the SNP with the largest NSFP score at each locus

Table 3

Genetic loci with successful signal replication in the COPDGene-AA-Severe cohort.

CHR	GENE[1]	SNP[2]	NSFP	$P_{\text{NHW-SEVERE}}$	$P_{\text{AA-SEVERE}}$	$P_{\text{META-RANDOM}}$[3]
4q31	*HHIP*	rs6837671	0.7810	9.73E-07	6.18E-03	1.98E-08
4q22	*FAM13A*	rs13105210	0.8798	2.79E-07	1.19E-02	1.74E-08
14q32	*RIN3*	rs754388	0.7472	1.02E-05	2.36E-02	7.09E-07
15q25	*CHRNA3*	rs12914385	0.9903	4.97E-11	2.20E-03	4.18E-13
20q11	*POFUT1*	rs11905172	0.7112	1.44E-05	2.02E-02	1.15E-06

[1] Only one gene at each locus is listed.

[2] The SNP with the lowest p-value in the random-effect meta-analysis is listed.

[3] p-values in the random-effect meta-analysis of NHW-severe and AA-severe datasets