ORIGINAL ARTICLE

Identification of FGF7 as a novel susceptibility locus for chronic obstructive pulmonary disease

John M Brehm,[1] Koichi Hagiwara,[2] Yohannes Tesfaigzi,[3] Shannon Bruse,[3] Thomas J Mariani,[4] Soumyaroop Bhattacharya,[4] Nadia Boutaoui,[1] John P Ziniti,[5] Manuel E Soto-Quiros,[6] Lydiana Avila,[6] Michael H Cho,[5,7,8] Blanca Himes,[5] Augusto A Litonjua,[5,7,8,9] Francine Jacobson,[10] Per Bakke,[11] Amund Gulsvik,[11] Wayne H Anderson,[12] David A Lomas,[13] Erick Forno,[14] Soma Datta,[5] Edwin K Silverman,[5,7,8,15] Juan C Celedón[1]

► Additional materials are published online only. To view these files please visit the journal online (http://thorax.bmj.com).

For numbered affiliations see end of article.

Correspondence to
Dr Juan C Celedón, Division of Pediatric Pulmonary Medicine, Allergy and Immunology, Children's Hospital of Pittsburgh of UPMC, 4401 Penn Avenue, Pittsburgh, PA 15224, USA; juan.celedon@chp.edu

Received 11 February 2011
Accepted 5 August 2011
Published Online First
15 September 2011

ABSTRACT

Rationale Traditional genome-wide association studies (GWASs) of large cohorts of subjects with chronic obstructive pulmonary disease (COPD) have successfully identified novel candidate genes, but several other plausible loci do not meet strict criteria for genome-wide significance after correction for multiple testing.

Objectives The authors hypothesise that by applying unbiased weights derived from unique populations we can identify additional COPD susceptibility loci.

Methods The authors performed a homozygosity haplotype analysis on a group of subjects with and without COPD to identify regions of conserved homozygosity haplotype (RCHHs). Weights were constructed based on the frequency of these RCHHs in case versus controls, and used to adjust the p values from a large collaborative GWAS of COPD.

Results The authors identified 2318 RCHHs, of which 576 were significantly ($p<0.05$) over-represented in cases. After applying the weights constructed from these regions to a collaborative GWAS of COPD, the authors identified two single nucleotide polymorphisms (SNPs) in a novel gene (fibroblast growth factor-7 (*FGF7*)) that gained genome-wide significance by the false discovery rate method. In a follow-up analysis, both SNPs (rs12591300 and rs4480740) were significantly associated with COPD in an independent population (combined p values of 7.9E−7 and 2.8E−6, respectively). In another independent population, increased lung tissue *FGF7* expression was associated with worse measures of lung function.

Conclusion Weights constructed from a homozygosity haplotype analysis of an isolated population successfully identify novel genetic associations from a GWAS on a separate population. This method can be used to identify promising candidate genes that fail to meet strict correction for multiple testing.

Key messages

What is the key question?

► Can information from isolated populations improve our ability to detect novel genetic variants in genome-wide association studies (GWASs)?

What is the bottom line?

► We identified statistically significant polymorphisms in a novel chronic obstructive pulmonary disease (COPD) gene (*FGF7*), which we replicated in an independent population.

Why read on?

► We demonstrate the use of a novel method (homozygosity haplotype analysis) for identifying genomic regions that are inherited from a common ancestor, and use this information to weight a GWAS of COPD to identify novel genetic variants that are associated with increased risk of disease.

INTRODUCTION

Traditional genome-wide association studies (GWASs) have identified novel susceptibility loci for complex diseases such as chronic obstructive pulmonary disease (COPD).[1–3] Because the effect size of most common disease-susceptibility variants is modest, GWASs of complex diseases require large sample sizes to achieve statistically significant results after correction for multiple testing. Weighting the results of GWASs according to prior information (eg, from linkage studies) may significantly improve the power to detect associations that do not meet genome-wide (GW) significance.[4]

Homozygosity mapping is a promising technique to identifying regions of the genome that are more likely to contain disease-susceptibility loci. Although initially developed to identify rare susceptibility mutations for monogenic traits in families,[5] homozygosity mapping has recently been successfully applied to the study of complex diseases.[6 7] While techniques vary, the concept underlying all homozygosity haplotype (HH) methods is that regions of homozygosity are more likely to contain disease-susceptibility loci in affected subjects than in unaffected individuals.[8]

Using high-density single nucleotide polymorphism (SNP) arrays, Miyazawa *et al* developed a novel variation of homozygosity mapping that tests whether multiple subjects share the same genotype among homozygous SNPs, and then constructed a region of conserved homozygosity

haplotype (RCHH) that reflects the transmission of the haplotype from a founder population. In theoretical simulations, this method was shown to be a viable method to detect disease-susceptibility loci in recently admixed populations.[9] We hypothesised that application of this method to a genetic isolate in Costa Rica would result in detection of an over-representation of regions of conserved homozygosity in subjects affected with COPD compared with unaffected subjects. In this report, we first identify regions of conserved homozygosity in Costa Ricans and then show that weights derived from these regions can be applied to GWASs in non-isolated populations to identify novel disease-susceptibility loci for COPD. Using this approach, we identify a novel COPD candidate gene (fibroblast growth factor-7 (*FGF7*)).

MATERIALS AND METHODS

Study population

The primary study population consisted of 58 subjects with COPD (cases) and 57 subjects without COPD (controls) in the Genetic Epidemiology of COPD in Costa Rica study. Cases were recruited from patients attending four adult hospitals in San José (Costa Rica) and their affiliated clinics, and through newspaper advertisements. Control subjects were recruited from individuals attending a smoking-cessation clinic at the Institute for Pharmaco-dependency in San José, and through newspaper advertisements. To ensure their descent from the founder population of the Central Valley of Costa Rica (which is predominantly of Spanish and Native American ancestry), all participants were required to have at least six great-grandparents born in the Central Valley. Additional inclusion criteria for cases were ages 21–71 years, physician-diagnosed COPD, ≥10 pack-years of cigarette smoking, a forced expiratory volume in one second (FEV_1) ≤65% predicted and an FEV_1/forced vital capacity (FVC) ratio of ≤70% after bronchodilator administration (180 μg of albuterol by metered dose inhaler). Controls were recruited on the basis of the same criteria for age and smoking history, but they had to have no physician-diagnosed COPD and normal spirometry. Exclusion criteria for cases and controls included history of chronic pre-existing chronic lung disease (eg, bronchiectasis) and severe α-1-antitrypsin deficiency (for cases), based on molecular phenotyping. The baseline characteristics of this cohort are in listed in the online supplementary table 2.

Written consent was obtained from participating subjects. The study was approved by the institutional review boards of the Hospital Nacional de Niños (San José, Costa Rica), Partners Healthcare System (Boston, Massachusetts, USA), and participating National Emphysema Treatment Trial (NETT), Evaluation of COPD Longitudinally to Identify Predictive Surrogate Endpoints (ECLIPSE) and Norway centres.

Genotyping of Costa Rican cohort

High-density SNP genotyping was performed using the Illumina Quad 610 platform at the Channing Laboratory, Boston, Massachusetts, USA. Cases and controls were randomly distributed among batches, and each batch contained a replicate sample. All subjects had an SNP call rate >95%. After quality control measures (see online supplementary table 1), a total of 558 929 SNPs were acceptable for analysis.

Collaborative COPD cohorts for the primary GWAS

Three populations with a total of 2940 cases and 1380 controls were used for the primary GWAS: (1) subjects in a case–control study of COPD in Norway (838 cases and 791 controls)[3]; (2) subjects in the NETT (366 cases) and the Normative Aging Study (414 controls)[10 11] and (3) 1736 cases and 175 controls from the multicentre ECLIPSE study.[12] All controls were current or former smokers with normal spirometry, and all cases with COPD had moderate to very severe disease according to the Global Initiative for Chronic Obstructive Lung Disease classification.[13]

Lovelace Smokers Cohort

The top SNPs in novel genes were replicated in a cohort of 1845 smoking adults in New Mexico, 424 (23%) of whom were classified with COPD based on an FEV_1/FVC ratio below the fifth percentile of the predicted value, also referred to as the lower limit of normal.[14] Of the 1845 participants, 1411 (77%) were Caucasian and 313 (17%) were Hispanic. The protocols for subject recruitment and data collection for the Lovelace Smokers Cohort have been previously described in detail.[15] The two SNPs (rs12591300 and rs4480740) were genotyped by allelic discrimination using Taqman assay (Applied Biosystems, Foster City, California, USA). The case–control association analysis was first performed in all subjects, and then separately in Caucasians and Hispanics. All analyses were adjusted for age, gender and pack-years of cigarette smoking; the analysis of all subjects was additionally adjusted for self-declared ethnicity.

Gene expression analysis

For the top novel candidate genes, we examined the correlation of gene expression in lung tissue with COPD intermediate phenotypes (FEV_1 and FEV_1/FVC ratio) in a previously published COPD biomarker discovery study.[16] This cohort consists of 56 subjects with varying degrees of obstruction who underwent lung resection for a solitary pulmonary nodule. RNA expression profiling was completed using the Affymetrix U133 Plus 2.0 array, as previously described.[16] Expression correlation with quantitative phenotypes was conducted as previously described.[16]

Statistical analysis

Construction of RCHHs

RCHHs were identified using the method described by Miyazawa *et al*.[9] In brief, for any given individual all heterozygous SNPs were ignored and the SNP location was scored with the value of the allele for that subject. Subjects are compared only across SNPs that are scored. RCHHs are defined by runs of SNPs that share the same allele at the homozygous locations across multiple subjects, ignoring heterozygous SNPs. The size of the shared segments between any two individuals was set at 3.0 cM (roughly and approximately three million base pairs), which in theoretical work conducted by Miyazawa *et al*[9] reduced the false positive and false negative rates of discovery. A theoretical ancestral segment was then constructed from the largest subgroup of subjects sharing a particular RCHH (see online supplementary figure 1). While any two subjects must have at least 3.0 cM of sharing, the size may be much smaller when comparing across multiple subjects (online supplementary figure 2). If more than one ancestral region is identified at a particular chromosomal location, the region shared by the most number of subjects is used (online supplementary figure 3). The total number of cases and controls sharing this ancestral allele is used to calculate a p value based on a standard normal distribution.

For the primary analysis of the collaborative COPD cohort, logistic regression analysis was performed under an additive genetic model for each SNP, adjusting for age, pack-years of smoking and the first 16 principal components (to adjust for population stratification). The p values from all RCHHs identified in Costa Rica were then used to construct a cumulative weight for each SNP from the recent GWAS of COPD in the combined cohort of Norway, ECLIPSE and NETT–Normative

Aging Study using the method developed by Roeder *et al*.[4] Briefly, the weighting method utilises prior information (in this case, the p value representing the degree of over-representation of a region of the genome in cases versus controls) to upweight or downweight p values from an association study (in this case, the GWAS of COPD in the collaborative cohort). In order to maintain an overall α level of 0.05, the assigned weights across the genome average to 1. For this study, SNPs that did not fall inside of an RCHH (and therefore did not have a p value) were assigned a p value of 1 (and therefore a weight approaching zero). This is a more conservative approach than excluding these SNPs from consideration. The method then calculates a false-discovery rate (FDR) using the method described by Benjamini and Hochberg[17] to correct for multiple testing.

The RCHHs were created and compared with HHAnalysis (available at http://www.hhanalysis.com). Association analysis was performed using PLINK V.1.07 (http://pngu.mgh.harvard.edu/purcell/plink). The weighting procedure was performed using software developed by Roeder *et al*[4] (http://wpicr.wpic.pitt.edu/wpiccompgen/). All other statistical analysis was performed using R V.2.9.0 (http://www.R-project.org).

RESULTS

Identification of RCHHs in Costa Rica and construction of weights

In total, 2318 RCHHs were identified in the Costa Rican cohort. Of these 2318 regions, 576 were significantly ($p<0.05$) over-represented in cases compared with controls; none of the regions were significantly more frequent in controls than cases. The median size of the significant regions was 105 kb, and the largest was 7.2 Mb. Online supplementary table 3 shows the top 20 p values representing 100 RCHHs in Costa Rica.

Each SNP in the combined collaborative COPD cohort was then mapped to an RCHH and assigned the p value of the whole region. SNPs that did not map to an RCHH were assigned a p value of 1. The mapped p values across all genotyped SNPs were then used to create weights using a cumulative distribution function. The algorithm is constructed so that the mean weight across all SNPs is 1: some SNPs are upweighted and a much larger fraction is downweighted. The nominal p value is divided by the weight to obtain the weighted p value.

Application of weights to the COPD GWAS

We applied the weights derived from the HH analysis above to reanalyse GW genotypic data in a cohort of subjects of European descent that was previously employed for a traditional GWAS of COPD. After weighting, 14 SNPs were significant at an FDR-corrected α of 0.05. The top five SNPs from the unweighted GWAS retained their original ranks, but several SNPs that did not achieve GW significance in the traditional GW association analysis became more statistically significant and moved higher in the list (table 1). Of these SNPs, those in the gene for *FAM13A* were identified in the original analysis of the GWAS,[1] and SNPs in *IREB2*[18] and *CHRNA3*[3] have been implicated in COPD affection status in prior candidate-gene and GWASs. Two of the other SNPs lie in two novel candidate genes for COPD, *FGF7* and proteasome subunit, α-type, 4 (*PSMA4*) (figure 1**)**. The RCHH in Costa Rica that contains *FGF7* was present in seven cases and no controls, and the RCHH containing *PSMA4* was present in five cases and no controls.

The regions containing the genes *CHRNA3* and *IREB2* were also over-represented in cases compared with controls ($p<0.05$), and after weighting they were GW significant by FDR. While there was an RCHH containing *FAM13A* identified in the Costa Rican cohort, it was only seen in one case and no controls.

Replication in Lovelace Smokers Cohort

The top two SNPs in or near *FGF7* were genotyped in the 1845 smoking adults in the Lovelace Smokers Cohort. The minor alleles of both SNPs conferred increased odds for COPD in the whole population in the same direction as the original collaborative COPD cohort (table 2). Among the Hispanic subgroup, the effect size was larger and in the same direction for both SNPs, but only rs12591300 showed a significant association with COPD affection status.

Gene expression analysis

Our previous studies indicate that gene expression patterns associated with quantitative, intermediate COPD phenotypes are most informative for the discovery of disease-associated genes.[16 18 19] We examined disease-associated expression patterns for our novel candidate genes in a previously published GW expression data set from 56 subjects with varying degrees of airflow obstruction (assessed by spirometric measures of lung function (FEV_1 and FEV_1/FVC ratio)).[16] Expression of *FGF7* (as defined by multiple and independent probe sets) was significantly negatively correlated with both FEV_1 (nominal p value <0.01) and FEV_1/FVC ratio (nominal p value <0.01), indicating increased expression associated with increased disease severity. Expression in COPD subjects was increased compared with control subjects, but the difference was not statistically significant. *PSMA4* expression was not correlated with lung function and was not differentially expressed in cases versus controls.

DISCUSSION

While successful in identifying novel candidate genes, GWASs of complex traits are unlikely to identify all potential common disease-susceptibility variants because of limited power if strict criteria for GW significance are applied. In the absence of a very large sample size, novel methods are needed to identify disease-susceptibility variants not meeting GW significance. We identified RCHHs for COPD in a GW case–control study in Costa Rica. After applying a weighting method based on the degree of significance of these regions to a GWAS of COPD cases and controls of European descent, we identified two SNPs in a novel candidate gene for COPD (*FGF7*) and demonstrated that several SNPs in the previously identified candidate genes *IREB2* and *CHRNA3* met GW criteria for statistical significance. An SNP in another novel gene (*PSMA4*) was GW significant after weighting. However, expression of *PSMA4* in the lung was not associated with COPD phenotypes, and thus the observed association is likely due to linkage disequilibrium with the nearby genes *CHRNA3* and *IREB2*. We then replicated the two *FGF7* SNPs in an independent cohort of smoking adults, and showed that they are both significantly associated in the same direction with COPD. Notably, the effect sizes in Hispanics are larger than in the overall cohort, suggesting that these alleles confer greater risk in this population. This Hispanic population in New Mexico has a similar proportion of European and Native American ancestry as the Costa Rican cohort,[20 21] so another likely possibility is that patterns of linkage disequilibrium may be different between Hispanics and Caucasians in this genomic region, and that these SNPs are tagging a haplotype or functional SNP in the Hispanic subjects. Additionally, there was a trend towards increased lung tissue expression of *FGF7* in an independent cohort of COPD subjects, in whom there was a significant negative correlation between *FGF7* expression and FEV_1 and FEV_1/FVC ratio.

Table 1 FDR significant* results from weighted GWAS

SNP	Location (Chr: BP in hg18)	A1	OR	Original p value†	Original rank	Weighted p value‡	Gene (distance from gene)	Norway p value§	NETT–NAS p value	Eclipse p value	Costa Rica RCHH		
											Cases (%)	Controls (%)	RCHH p value
rs1903003	4:90105320	C	0.75	7.18E−8	1	6.87E−8	FAM13A(0)	4.3E−4	1.4E−3	9.1E−3	1 (2%)	0 (0%)	0.2126
rs7671167	4:90103002	C	0.76	8.59E−8	2	8.22E−8	FAM13A(0)	7.9E−4	2.7E−4	7.8E−3	1 (2%)	0 (0%)	0.2126
rs1062980	15:76579582	C	0.76	4.81E−7	3	9.53E−8	IREB2(0)	9.9E−3	1.0E−2	3.6E−2	5 (9%)	0 (0%)	0.0164
rs13180	15:76576543	C	0.76	5.01E−7	4	9.93E−8	IREB2(0)	7.9E−3	1.6E−2	4.3E−2	5 (9%)	0 (0%)	0.0164
rs8034191	15:76593078	C	1.32	5.37E−7	5	1.06E−7	IREB2 (+12.22 kb)	1.5E−4	8.7E−3	8.2E−1	5 (9%)	0 (0%)	0.0164
rs12914385	15:76685778	T	1.29	1.42E−6	9	2.81E−7	CHRNA3(0)	1.4E−3	9.8E−3	9.5E−1	5 (9%)	0 (0%)	0.0164
rs1051730	15:76681394	A	1.29	2.80E−6	14	5.54E−7	CHRNA3(0)	4.3E−4	2.1E−2	8.4E−1	5 (9%)	0 (0%)	0.0164
rs17404727	15:47791375	C	1.28	4.71E−6	15	7.17E−7		1.9E−2	2.2E−2	3.4E−2	7 (12%)	0 (0%)	0.0049
rs996414	9:26570067	G	0.76	1.80E−6	11	8.76E−7		6.2E−4	8.8E−1	2.6E−1	2 (3%)	0 (0%)	0.1063
rs4480740	15:47543134	A	1.27	6.75E−6	17	1.03E−6	FGF7(0)	4.0E−2	2.5E−2	2.1E−2	7 (12%)	0 (0%)	0.0049
rs12591300	15:47492033	A	1.27	8.78E−6	21	1.34E−6	FGF7 (−10.72 kb)	3.9E−2	8.3E−2	2.5E−2	7 (12%)	0 (0%)	0.0049
rs2656069	15:76532762	C	0.75	6.82E−6	18	1.35E−6	IREB2(0)	1.6E−1	2.6E−3	1.0E−2	5 (9%)	0 (0%)	0.0164
rs2036534	15:76614003	C	0.75	6.98E−6	19	1.38E−6	PSMA4 (−5.798 kb)	5.8E−2	7.4E−3	9.4E−2	5 (9%)	0 (0%)	0.0164
rs2869967	4:90088355	C	1.29	1.48E−6	10	1.41E−6	FAM13A1 (0)	4.7E−4	7.6E−3	4.4E−3	1 (2%)	0 (0%)	0.2126

*An FDR-corrected p value of 1.43E−6 was used as the cut-off for genome-wide significance.
†Results previously published by Cho *et al*.[1]
‡The weighted p value is the original p value divided by the weight constructed from the RCHH (not shown).
§p Values for individual cohorts are the original, unweighted p values.
COPD, chronic obstructive pulmonary disease; ECLIPSE, Evaluation of COPD Longitudinally to Identify Predictive Surrogate Endpoints; FDR, false-discovery rate; FGF7, fibroblast growth factor-7; GWAS, genome-wide association study; NAS, Normative Aging Study; NETT, National Emphysema Treatment Trial; PSMA4, proteasome subunit, α-type, 4; RCHH, region of conserved homozygosity haplotype; SNP, single nucleotide polymorphism.

There are several plausible methods for weighting chromosomal regions in GWAS, including upweighting previously identified candidate genes, coding variants, exons and promoter regions. However, these weighting strategies work counter to one of the strengths of a GWAS: its hypothesis-free nature. Using HHs as a weighting method avoids the pitfall of these

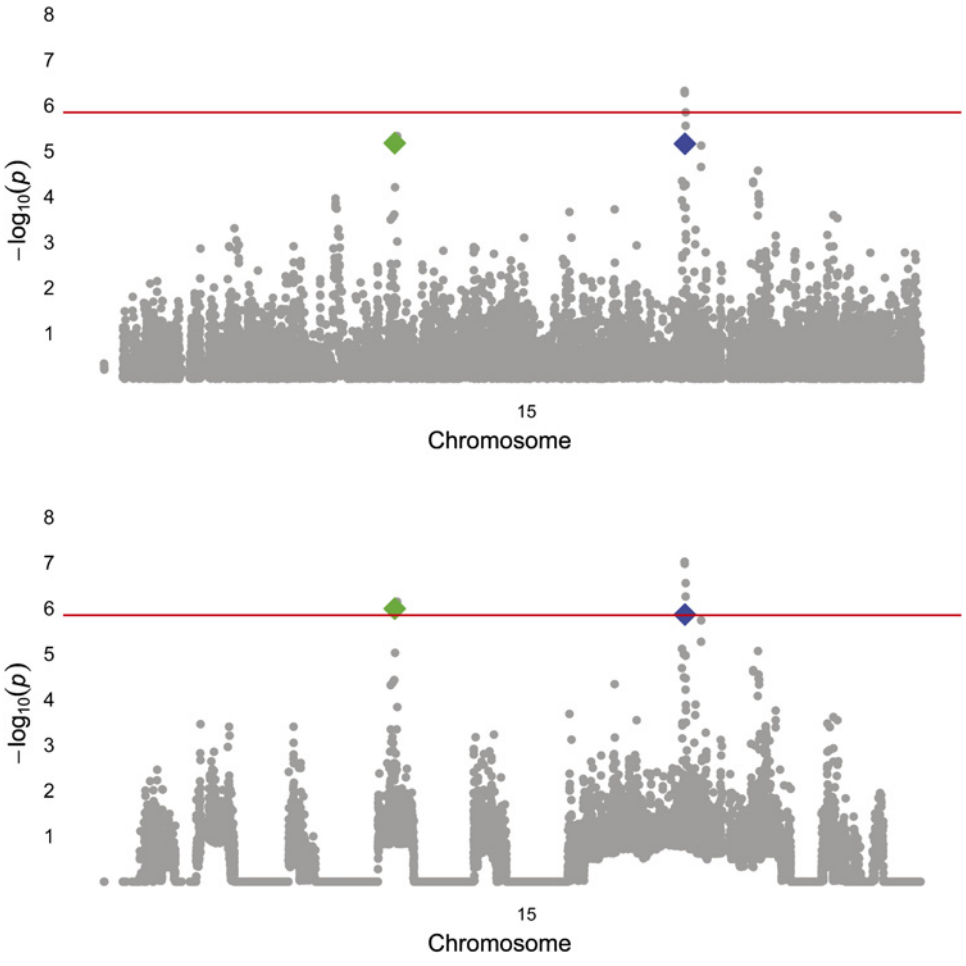


Figure 1 Manhattan plot of chromosome 15, before (top) and after weighting. rs4480740 (Green) is in the gene FGF7 and rs2036534 (blue) is in the promoter of PSMA4. The red line indicates the FDR corrected α level for genome-wide significance. FDR, false-discovery rate; FGF7, fibroblast growth factor-7; PSMA4, proteasome subunit, α-type, 4.

Table 2 Combined p values for replication of FGF7 SNPs

SNP	COPD consortium OR (two-sided p value)	Lovelace (all subjects) OR (one-sided p value)	Lovelace (Hispanics) OR (one-sided p value)	Combined one-sided p value (all subjects)	Combined one-sided p value (Hispanics only)*
rs12591300	1.27 (8.78E−6)	1.21 (0.01)	1.5 (0.06)	7.9E−7	3.9E−6
rs4480740	1.27 (6.75E−6)	1.15 (0.05)	1.64 (0.025)	2.8E−6	1.5E−6

*Fisher's combined p value using original two-sided p values.
COPD, chronic obstructive pulmonary disease; FGF7, fibroblast growth factor-7; SNP, single nucleotide polymorphism.

other weighting strategies because they are constructed using a hypothesis-free method, so the weights are unbiased with respect to prior knowledge.

One of the main strengths of our study is that it shows the power of using HH analysis in an isolated population to investigate common diseases. While our sample size was small, Miyazawa *et al*[9] have previously shown in simulated data that HH analysis has the ability to identify the region containing an SNP inherited identity-by-descent from a distant common ancestor using only 45 cases and 45 controls. In our own data, we were able to show that the previously identified candidate genes *IREB2* and *CHRNA3* fall within an RCHH that is significantly over-represented in subjects with COPD. When combined with results from a weighted GWAS in an independent cohort with adequate sample size, we were able to show that variants in these genes are significant after correction for multiple testing.

Two novel genes are contained within significant regions of conserved homozygosity, and after weighting they are significant by FDR correction. The first, *FGF7*, was identified in cultured human embryonic lung fibrobasts,[22] and plays a role in promoting wound healing[23] and protecting airway epithelium from oxidant injury in mice.[24] One of the SNPs identified in this study (rs4480740) is in an intron of *FGF7*, and the other (rs12591300) lies immediately upstream of *FGF7* in an intron of hypothetical protein LOC196951. In a GWAS of FEV_1 in the British 1958 Birth Cohort, five out of the nine SNPs genotyped in FGF7 were significantly ($p<0.05$) associated with differences in lung function, although not the two SNPs identified in this study.[25] FGF7 has been shown to protect against oxidative stress response specifically in the lung epithelium,[24] so increases in expression associated with disease progression may indicate a greater burden of injury. A limitation of our study is the lack of experimental evidence for an effect(s) of the SNPs identified in *FGF7* on gene expression. We hypothesise that these SNPs cause decreased expression of *FGF7*, which could affect antioxidant mechanisms protecting against detrimental effects of cigarette smoking on the lung. Alternatively, *FGF7* may play a role in disease susceptibility through its role in epithelial development during embryogenesis by influencing epithelial responses to cigarette smoke. Since it is unclear whether increased *FGF7* expression is a marker of exposure to oxidant injury or a cause of epithelial damage, further work must be done to characterise the role of these SNPs on *FGF7* expression.

The HHAnalysis algorithm works best under certain assumptions, namely that (1) the risk alleles were introduced into the population from a population of common ancestors within the last several hundred years, (2) the target population is genetically isolated, (3) the number of common ancestors introducing the risk allele is small and that (4) the risk of the disease allele is moderate to high. Violations of these assumptions reduce the theoretical expected size of the RCHH and/or the association of the RCHH with disease, which reduces the power of the algorithm to detect them. Genetic and historical data for the population of the Central Valley of Costa Rica suggest that the first three assumptions are met. As in most association studies of complex disease, the effect size of a risk allele is likely small to moderate at most, and we expect that this has somewhat reduced our power.

Whereas other homozygosity mapping methods are primarily designed to detect recessive alleles, the HHAnalysis method instead uses homozygosity to identify ancestral regions inherited from a common ancestor. These regions from a common ancestor can harbour risk alleles that operate under recessive, dominant or additive models. However, the HHAnalysis algorithm would also detect copy number variation that results in the deletion of a single allele. While this may explain a fraction of the regions identified, the top novel SNPs identified in *FGF7* do not fall within known regions of copy number variation according to the Database of Genomic Variants.[26]

In summary, we have shown that weights obtained from HH analysis in an isolated population can improve the power to detect novel variants in GWAS in non-isolates. In addition to confirming results for previously identified variants in *IREB2* and *CHRNA3*, we have identified variants in a novel candidate gene (*FGF7*) for COPD. The validity of this gene is supported by replication in an independent cohort of smoking adults, and expression data showing consistent and significant patterns associated with COPD intermediate lung function phenotypes. Further analysis of these genes in the Costa Rican cohort and functional studies should yield insights into the causative SNPs or haplotypes that underlie the associations identified in this study.

Author affiliations
[1]Division of Pediatric Pulmonary Medicine, Allergy, and Immunology, Children's Hospital of Pittsburgh of UPMC, Pittsburgh, Pennsylvania, USA
[2]Saitama Medical University Hospital and Institute, Saitama, Japan
[3]Lovelace Respiratory Research Institute, Albuquerque, New Mexico, USA
[4]University of Rochester Medical Center, Rochester, New York, USA
[5]Channing Laboratory, Brigham and Women's Hospital, Boston, Massachusetts, USA
[6]Division of Pediatric Pulmonology, Hospital Nacional de Niños, San José, Costa Rica
[7]Division of Pulmonary/Critical Care Medicine, Brigham and Women's Hospital, Boston, Massachusetts, USA
[8]Department of Medicine, Brigham and Women's Hospital, Boston, Massachusetts, USA
[9]Department of Medicine, Harvard Medical School, Boston, Massachusetts, USA
[10]Department of Radiology, Brigham and Women's Hospital, Boston, Massachusetts, USA
[11]Haukeland University Hospital and Institute of Medicine, University of Bergen, Bergen, Norway
[12]GlaxoSmithKline Research and Development, Research Triangle Park, North Carolina, USA
[13]Cambridge Institute for Medical Research, University of Cambridge, Cambridge, UK
[14]Division of Pediatric Pulmonology, Department of Pediatrics, University of Miami, Miami, Florida, USA
[15]Center for Genomic Medicine, Brigham and Women's Hospital, Boston, Massachusetts, USA

Funding The Genetic Epidemiology of COPD in Costa Rica is supported by grant R01HL073373 from the National Heart, Lung, and Blood Institute. The National Emphysema Treatment Trial (NETT) is supported by contracts with the National Heart, Lung, and Blood Institute (N01HR76101, N01HR76102, N01HR76103, N01HR76104, N01HR76105, N01HR76106, N01HR76107, N01HR76108, N01HR76109, N01HR76110, N01HR76111, N01HR76112, N01HR76113, N01HR76114, N01HR76115, N01HR76116, N01HR76118, N01HR76119), the Centers for Medicare and Medicaid Services (CMS) and the Agency for Healthcare Research and Quality (AHRQ). The Norway cohort and the ECLIPSE study (http://clinicaltrials.gov identifier NCT00292552; GSK Code SCO104960) are funded by GlaxoSmithKline. The Lovelace Smokers Cohort is supported by funding from the State of New Mexico (appropriation from the Tobacco Settlement Fund) and by grant RO1 ES015482 from the National

Institute of Environmental Health Sciences. We acknowledge use of genotype data from the British 1958 Birth Cohort DNA collection, funded by the Medical Research Council grant G0000934 and the Wellcome Trust grant 068545/Z/02.

Competing interests None.

Ethics approval Institutional Review Board of University of Pittsburgh, Partners Health Care (Boston), participating NETT centres, Boston VA, Norway, Costa Rica, and Lovelace Respiratory Institute.

Contributors Manuscript preparation: JMB, JCC; data analysis and study design: JMB, YT, SB, TJM, SB, NB, JPZ, MES-O, LA, MHC, BH, AAL, FJ, EF, SD, EKS, JCC; data collection: YT, TJM, MES, LA, AAL, PB, AG, WHA, DAL, EKS, JCC; statistical analysis: JMB, KH, MHC.

Provenance and peer review Not commissioned; externally peer reviewed.

REFERENCES

1. **Cho MH,** Boutaoui N, Klanderman BJ, *et al*. Variants in FAM13A are associated with chronic obstructive pulmonary disease. *Nat Genet* 2010;**42**:200–2.
2. **Hirschhorn JN.** Genomewide association studies—illuminating biologic pathways. *N Engl J Med* 2009;**360**:1699–701.
3. **Pillai SG,** Ge D, Zhu G, *et al*. A genome-wide association study in chronic obstructive pulmonary disease (COPD): identification of two major susceptibility loci. *PLoS Genet* 2009;**5**:e1000421.
4. **Roeder K,** Bacanu SA, Wasserman L, *et al*. Using linkage genome scans to improve power of association in genome scans. *Am J Hum Genet* 2006;**78**:243–52.
5. **Lander ES,** Botstein D. Homozygosity mapping: a way to map human recessive traits with the DNA of inbred children. *Science* 1987;**236**:1567–70.
6. **Lencz T,** Lambert C, DeRosse P, *et al*. Runs of homozygosity reveal highly penetrant recessive loci in schizophrenia. *Proc Natl Acad Sci U S A* 2007;**104**:19942–7.
7. **Morrow EM,** Yoo SY, Flavell SW, *et al*. Identifying autism loci and genes by tracing recent shared ancestry. *Science* 2008;**321**:218–23.
8. **Reich DE,** Schaffner SF, Daly MJ, *et al*. Human genome sequence variation and the influence of gene history, mutation and recombination. *Nat Genet* 2002;**32**:135–42.
9. **Miyazawa H,** Kato M, Awata T, *et al*. Homozygosity haplotype allows a genomewide search for the autosomal segments shared among patients. *Am J Hum Genet* 2007;**80**:1090–102.
10. **Bell B,** Rose CL, Damon A. The Normative Aging Study: an interdisciplinary and longitudinal study of health and aging. *Aging Hum Dev* 1972;**3**:5–17.
11. **Fishman A,** Martinez F, Naunheim K, *et al*. A randomized trial comparing lung-volume-reduction surgery with medical therapy for severe emphysema. *N Engl J Med* 2003;**348**:2059–73.
12. **Vestbo J,** Anderson W, Coxson HO, *et al*. Evaluation of COPD Longitudinally to Identify Predictive Surrogate End-points (ECLIPSE). *Eur Respir J* 2008;**31**:869–73.
13. **Pauwels RA,** Buist AS, Calverley PM, *et al*. Global strategy for the diagnosis, management, and prevention of chronic obstructive pulmonary disease. NHLBI/WHO Global Initiative for Chronic Obstructive Lung Disease (GOLD) Workshop summary. *Am J Respir Crit Care Med* 2001;**163**:1256–76.
14. **Pellegrino R,** Viegi G, Brusasco V, *et al*. Interpretative strategies for lung function tests. *Eur Respir J* 2005;**26**:948–68.
15. **Sood A,** Stidley CA, Picchi MA, *et al*. Difference in airflow obstruction between Hispanic and non-Hispanic White female smokers. *COPD* 2008;**5**:274–81.
16. **Bhattacharya S,** Srisuma S, Demeo DL, *et al*. Molecular biomarkers for quantitative and discrete COPD phenotypes. *Am J Respir Cell Mol Biol* 2009;**40**:359–67.
17. **Benjamini Y,** Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J Roy Stat Soc B Stat Meth* 1995;**57**:289–300.
18. **DeMeo DL,** Mariani T, Bhattacharya S, *et al*. Integration of genomic and genetic approaches implicates IREB2 as a COPD susceptibility gene. *Am J Hum Genet* 2009;**85**:493–502.
19. **Demeo DL,** Mariani TJ, Lange C, *et al*. The SERPINE2 gene is associated with chronic obstructive pulmonary disease. *Am J Hum Genet* 2006;**78**:253–64.
20. **Klimentidis YC,** Miller GF, Shriver MD. Genetic admixture, self-reported ethnicity, self-estimated admixture, and skin pigmentation among Hispanics and Native Americans. *Am J Phys Anthropol* 2009;**138**:375–83.
21. **Morera B,** Barrantes R, Marin-Rojas R. Gene admixture in the Costa Rican population. *Ann Hum Genet* 2003;**67**:71–80.
22. **Rubin JS,** Osada H, Finch PW, *et al*. Purification and characterization of a newly identified growth factor specific for epithelial cells. *Proc Natl Acad Sci U S A* 1989;**86**:802–6.
23. **Werner S,** Smola H, Liao X, *et al*. The function of KGF in morphogenesis of epithelium and reepithelialization of wounds. *Science* 1994;**266**:819–22.
24. **Ray P,** Devaux Y, Stolz DB, *et al*. Inducible expression of keratinocyte growth factor (KGF) in mice inhibits lung epithelial cell death induced by hyperoxia. *Proc Natl Acad Sci U S A* 2003;**100**:6098–103.
25. http://www.b58cgene.sgul.ac.uk/ (accessed 8 Jan 2010).
26. **Iafrate AJ,** Feuk L, Rivera MN, *et al*. Detection of large-scale variation in the human genome. *Nat Genet* 2004;**36**:949–51.

Journal club

The increasing importance of innate antimicrobials

The protein short palate, lung and nasal epithelium clone 1 (SPLUNC1) is found in airway epithelium. It is known to have reduced expression in chronic airways diseases and in smokers. This study aimed to demonstrate its antimicrobial properties against *Mycoplasma pneumoniae* infection. The investigators compared the inflammatory and antibacterial responses to *M pneumoniae* infection in transgenic mice deficient in expressing this protein with the responses in mice overexpressing the protein.

The overall results showed SPLUNC1 has antibacterial effects by inhibiting bacterial adherence proteins, thereby inhibiting *M pneumoniae* growth. Following *M pneumoniae* infection, a reduction in tissue inflammation and increase in neutrophil elastase production was seen in those mice with expression of SPLUNC1. Neutrophil elastase is important in relation to infection and was shown to reduce *M pneumoniae* growth when incubated with human sputum neutrophil elastase.

This study shows the potential antibacterial and immunomodulatory functions of SPLUNC1, which may help in the development of novel treatments for chronic airway diseases. In an ever-increasing climate of drug resistance, it emphasises the importance of focusing on host endogenous antimicrobial responses.

► **Fabienne G**, Di YP, Smith SK. SPLUNC1 promotes lung innate defense against mycoplasma pneumoniae in mice. *Am J Path* 2011;**178**:2159–67.

A Loughnan

Correspondence to A Loughnan, Homerton University Hospital, London, UK; alice.loughnan@homerton.nhs.uk

Published Online First 31 July 2011

Thorax 2011;**66**:1090. doi:10.1136/thoraxjnl-2011-200801

Identification of FGF7 as a novel susceptibility locus for chronic obstructive pulmonary disease

John M. Brehm,[1-3,5,7] Koichi Hagiwara,[9] Y Tesfaigzi Y,[15] S Bruse,[15] Thomas J. Mariani,[10] Soumyaroop Bhattacharya,[10] Nadia Boutaoui,[1] John P. Ziniti,[2] Manuel E. Soto-Quiros,[8] Lydiana Avila,[8] Michael H. Cho, [2,3,5] Blanca Himes,[2] Augusto A. Litonjua, [2,3,5,7] Francine Jacobson,[6] Per Bakke,[12] Amund Gulsvik,[12] Wayne H Anderson,[13] David A. Lomas,[14] Erick Forno, [11] Soma Datta,[2] Edwin K. Silverman,[2-5] and Juan C. Celedón[1,2-5,7]

[1]Division of Pediatric Pulmonary Medicine, Allergy, and Immunology, Children's Hospital of Pittsburgh of UPMC, Pittsburgh, PA; [2]Channing Laboratory, [3]Division of Pulmonary/Critical Care Medicine, [4]Center for Genomic Medicine, [5]Department of Medicine, and [6]Department of Radiology, Brigham and Women's Hospital, Boston, Massachusetts; [7]Department of Medicine, Harvard Medical School, Boston, Massachusetts; [8]Division of Pediatric Pulmonology, Hospital Nacional de Niños, San José, Costa Rica; [9]Saitama Medical University Hospital and Institute, Saitama, Japan; [10]University of Rochester Medical Center, Rochester, NY, [11]Division of Pediatric Pulmonology, Dept. of Pediatrics, University of Miami, Miami, FL, [12]Haukeland University Hospital and Institute of Medicine, University of Bergen, Bergen, Norway; [13]GlaxoSmithKline Research and Development, Research Triangle Park, North Carolina, USA; [14]Cambridge Institute for Medical Research, University of Cambridge, Cambridge, UK; [15]Lovelace Respiratory Research Institute, Albuquerque, NM

Supplementary Methods

Study Populations

Costa Rica: The primary study population consisted of 58 subjects with (cases) and 57 subjects without (controls) COPD in the Genetic Epidemiology of COPD in Costa Rica (GECCOR) study. Cases were recruited from patients attending four adult hospitals in San José (Costa Rica) and their affiliated clinics, and through newspaper ads. Control subjects were recruited from individuals attending a smoking-cessation clinic at the Institute for Pharmaco-dependency in San José, and through newspaper ads. All of these control subjects had no addictions other than to nicotine as determined by phone questionnaire. To ensure their descent from the founder population of the Central Valley of Costa Rica (which is predominantly of Spanish and Native American ancestry), all participants were required to have at least six great-grandparents born in the Central Valley of Costa Rica. Additional inclusion criteria for cases were age 21 to 71 years, physician-diagnosed COPD, ≥ 10 pack-years of cigarette smoking, and an FEV1 ≤ 65% predicted and an FEV1/FVC ratio of ≤ 70% after bronchodilator administration. Controls were recruited on the basis of the same criteria for age and smoking history, but they had to have no physician-diagnosed COPD and normal spirometry. Exclusion criteria for cases and controls included history of chronic pre-existing chronic lung disease (e.g., bronchiectasis). Pre-existing lung disease was determined by physician diagnosis, CT scans (41/58 cases), and subject questionnaire. Controls who had significant dyspnea, cough, or wheezing without a formal pulmonary diagnosis were also excluded.

Evaluation of COPD Longitudinally to Identify Predictive Surrogate End-Points (ECLIPSE; SCO104960, NCT00292552): Details of the ECLIPSE study have been described previously[1] Briefly, ECLIPSE is a 3-year longitudinal study of COPD. Cases were diagnosed by post-bronchodilator spirometry[2] as GOLD Stage II or greater (forced expiratory volume in 1 second (FEV1) < 80% predicted and FEV1/ forced vital capacity (FVC) < 0.7) and without severe alpha-1 antitrypsin deficiency; smoking controls had normal lung function (post-bronchodilator FEV1 >85% predicted and FEV1/FVC > 0.7). Both cases

and controls were limited to subjects of self-reported white ethnicity, between ages 40-75 and were required to have at least a 10 pack-year smoking history. As the study was designed to evaluate COPD-related endpoints, recruitment was weighted towards cases; a total of 1839 cases and 196 smoking controls were genotyped.

National Emphysema Treatment Trial (NETT): Details of the National Emphysema Treatment Trial, a clinical trial to evaluate lung volume reduction surgery, have been published previously[3]. NETT subjects were enrolled at 17 clinical centers based on severe airflow obstruction by spirometry[2] (FEV1 ≤ 45% predicted[4]) and emphysema on computed tomographic (CT) imaging of the chest. The NETT Genetics Ancillary Study contains a subset of the original cohort with blood available for genotyping. After providing written informed consent, NETT participants provided a blood sample for DNA extraction for genetic studies of COPD. The study was approved by the institutional review boards at participating NETT centers. A total of 382 self-reported white subjects without severe alpha-1 antitrypsin deficiency were included in this study.

Normative Aging Study (NAS): The Normative Aging Study is an ongoing longitudinal study of healthy men established in 1963 and conducted by the Veterans Administration (VA) [5]. Briefly, men from the greater Boston area, ages 21 to 80 years, enrolled in the study after an initial health screening determined that they were free of known chronic medical conditions. Since enrollment, the participants have undergone comprehensive clinical examinations at 5-year intervals for those < 52 years old and at 3-year intervals for those > 52 years old. Selection of controls for COPD genetic studies from this population has been described previously[6]; self-reported white control subjects had DNA available for genotyping, a history of at least 10 pack-years of cigarette smoking and no evidence for airflow obstruction by spirometry[2] at their most recent visit (FEV1 > 80% predicted[7] and FEV1/FVC > 90% predicted[4]. Anonymized data were used, as approved by the institutional review boards of Partners Healthcare System and the Boston VA. A total of 453 subjects meeting enrollment criteria were

included in this study.

Norway: Details on the Bergen, Norway case-control study have been described previously[8]. Subjects were recruited from Bergen, Norway, and in contrast to the NAS and ECLIPSE studies, were required to have > 2.5 pack years of smoking history. GOLD Stage II or greater COPD cases were diagnosed by post-bronchodilator spirometry2 (FEV1 < 80% predicted and FEV1/FVC < 0.7), while controls had normal spirometry (FEV1 > 80% predicted and FEV1/FVC > 0.7). Subjects with alpha-1 antitrypsin genotypes PI ZZ, ZNull, Null-Null or SZ were excluded. All subjects gave informed consent, and the appropriate institutional review boards approved the study. A total of 933 cases and 919 controls of self-identified white ethnicity were included in this study.

Lovelace Smokers Cohort: Details on the Lovelace Smokers Cohort have been published previously.[9] Our top two SNPs in *FGF7* were replicated using a Taqman assay (Applied Biosystems, Foster City, CA). All subjects gave informed consent, and the institutional review board of the Lovelace Respiratory Institute approved the study. COPD was classified based on an FEV1/FVC ratio below the 5th percentile of the predicted value, also referred to as the lower limit of normal (LLN).[10]

Genotyping and Quality Control

Genotyping of NETT,NAS,Norway, and ECLIPSE has been described in detail previously [11].
For the Costa Rica cohort, genotyping was performed on the Illumina Quad 610 platform (Illumina, Inc; San Diego,CA) at the Channing Laboratory, Brigham and Women's Hospital. Cases and controls were randomly allocated in batches, which included at least one replicate sample. First pass cleaning using two-channel intensity was performed using Beadstudio. Although subjects with <95% of markers genotyped, and SNPs with a call frequency < 95% were to be removed a priori, all genotyped subjects passed with an average call rate of 99.87%. The remaining markers were cleaned following Illumina guidelines.

(http://www.illumina.com/documents/products/technotes/technote_infinium_genotyping_data_analysis.pdf)

SNPs with a call frequency between 0.9 and 0.99 and a cluster separation below 0.3 were manually inspected, with a threshold set for each workspace below which no SNPs were deemed unambiguous. We additionally examined SNPs with a low ABR mean, ABT mean, and heterozygote excess. X, Y, and mitochondrial SNPs were excluded from analysis.

Subject and Marker Cleaning

For the Costa Rica cohort, BeadStudio workspaces were exported to ped file format and further quality control assessment was performed using Python (www.python.org) and R (www.r-project.org) scripts in conjunction with PLINK (1.05) [12]. Quality control of genotyping data was assessed by subject and by marker. Subject data were excluded after examining missingness, reproducibility and discordances, relatedness, sex, and inbreeding. Subjects with discordance > 1% in replicate genotyping were discarded. Relatedness was examined using rgGrr[13] and estimated IBD in PLINK, using a cutoff of 0.12513. Within each set of related subjects, unrelated individual(s) were chosen based on phenotypic criteria. In addition to testing for relatedness within each cohort, a test for relatedness between cohorts was performed prior to construction of the primary (merged) dataset. Sex assignment was based on X homozygosity estimates, with those > 0.8 as male, and < 0.2 as female. Discordant samples were removed. Inbreeding coefficients > |0.2| were removed. Finally, among samples run in replicate, the sample with the higher passing genotype rate was chosen.

Markers were excluded based on missingness, minor allele frequency, reproducibility, and Hardy-Weinberg equilibrium. Markers with missingness > 5% were excluded. A strict minor allele frequency cutoff was not specified; instead, markers that were monoallelic or singleton in each dataset were excluded. Within replicates, SNPs showing significant discordance (examining the distribution of discordances) were excluded. Markers with Hardy-Weinberg deviation – P value less than .01 – were also excluded. Differential missingness between cases and controls was assessed, but not used to exclude markers at this stage. A summary of the quality control measures is shown in Supplementary

Table 1.

Construction of RCHH

Regions of conserved homozygosity haplotype (RCHHs) were identified using the method described by Miyazawa et. al. [14]. In brief, for any given individual all heterozygous SNPs were ignored and the SNP location was scored with the value of the allele for that subject. Subjects are compared only across SNPs that are scored. RCHH are defined by runs of SNPs that share the same allele at the homozygous locations across multiple subjects, ignoring heterozygous SNPs. The size of the shared segments between any two individuals was set at 3.0 cM, which is a size designed to reduce the false positive and false negative rates of discovery. A theoretical ancestral segment is then constructed from the largest subgroup of subjects sharing a particular RCHH (see Supplementary Figure 1). While any two subjects much have at least 3.0 cM of sharing, the size may be much smaller when comparing across multiple subjects (Supplementary figure 2). The total number of cases and controls sharing this ancestral allele is used to calculate a P value based on a standard normal distribution.

Supplementary Acknowledgments

The authors would like to acknowledge the ECLIPSE steering and scientific committees and investigators (listed below). ECLIPSE Steering Committee: Harvey Coxson (Canada), Lisa Edwards (GlaxoSmithKline, USA), Katharine Knobil (Co-chair, GlaxoSmithKline, UK), David Lomas (UK), William MacNee (UK), Edwin Silverman (USA), Ruth Tal-Singer (GlaxoSmithKline, USA), Jørgen Vestbo (Co-chair, Denmark), Julie Yates (GlaxoSmithKline, USA).

ECLIPSE Scientific Committee: Alvar Agusti (Spain), Peter Calverley (UK), Bartolome Celli (USA), Courtney Crim (GlaxoSmithKline, USA), Bruce Miller (GlaxoSmithKline, UK), William MacNee (Chair, UK), Stephen Rennard (USA), Ruth Tal-Singer (GlaxoSmithKline, USA), Emiel Wouters (The Netherlands), Julie Yates (GlaxoSmithKline, USA).

ECLIPSE Investigators: Bulgaria: Yavor Ivanov, Pleven; Kosta Kostov, Sofia. Canada: Jean Bourbeau, Montreal, Que Mark Fitzgerald, Vancouver, BC; Paul Hernandez, Halifax, NS; Kieran Killian, Hamilton,

On; Robert Levy, Vancouver, BC; Francois Maltais, Montreal, Que; Denis O'Donnell, Kingston, On. Czech Republic: Jan Krepelka, Praha. Denmark: Jørgen Vestbo, Hvidovre. Netherlands: Emiel Wouters, Horn-Maastricht. New Zealand: Dean Quinn, Wellington. Norway: Per Bakke, Bergen. Slovenia: Mitja Kosnik, Golnik. Spain: Alvar Agusti, Jaume Sauleda, Palma de Mallorca. Ukraine: Yuri Feschenko, Kiev; Vladamir Gavrisyuk, Kiev; Lyudmila Yashina, Kiev; Nadezhda Monogarova, Donetsk. United Kingdom: Peter Calverley, Liverpool; David Lomas, Cambridge; William MacNee, Edinburgh; David Singh, Manchester; Jadwiga Wedzicha, London. United States of America: Antonio Anzueto, San Antonio, TX; Sidney Braman, Providence, RI; Richard Casaburi, Torrance CA; Bart Celli, Boston, MA; Glenn Giessel, Richmond, VA; Mark Gotfried, Phoenix, AZ; Gary Greenwald, Rancho Mirage, CA; Nicola Hanania, Houston, TX; Don Mahler, Lebanon, NH; Barry Make, Denver, CO; Stephen Rennard, Omaha, NE; Carolyn Rochester, New Haven, CT; Paul Scanlon, Rochester, MN; Dan Schuller, Omaha, NE; Frank Sciurba, Pittsburgh, PA; Amir Sharafkhaneh, Houston, TX; Thomas Siler, St. Charles, MO, Edwin Silverman, Boston, MA; Adam Wanner, Miami, FL; Robert Wise, Baltimore, MD; Richard ZuWallack, Hartford, CT.

NETT Study: We acknowledge the co-investigators in the NETT Genetics Ancillary Study including Joshua Benditt, Gerard Criner, Malcolm DeCamp, Philip Diaz, Mark Ginsburg, Larry Kaiser, Marcia Katz, Mark Krasna, Neil MacIntyre, Barry Make, Rob McKenna, Fernando Martinez, Zab Mosenifar, John Reilly, Andrew Ries, Paul Scanlon, Frank Sciurba, and James Utz.

Supplementary References

1. Vestbo J, Anderson W, Coxson HO, Crim C, Dawber F, Edwards L, et al. Evaluation of COPD Longitudinally to Identify Predictive Surrogate End-points (ECLIPSE). *Eur Respir J* 2008;31(4):869-73.
2. Standardization of Spirometry, 1994 Update. American Thoracic Society. *Am J Respir Crit Care Med* 1995;152(3):1107-36.
3. Fishman A, Martinez F, Naunheim K, Piantadosi S, Wise R, Ries A, et al. A randomized trial comparing lung-volume-reduction surgery with medical therapy for severe emphysema. *N Engl J Med* 2003;348(21):2059-73.
4. Crapo RO, Morris AH. Standardized single breath normal values for carbon monoxide diffusing capacity. *Am Rev Respir Dis* 1981;123(2):185-9.
5. Bell B, Rose CL, Damon A. The Normative Aging Study: an interdisciplinary and longitudinal study of health and aging. *Aging Hum Dev* 1972;3:5-17.
6. Hersh CP, Demeo DL, Lange C, Litonjua AA, Reilly JJ, Kwiatkowski D, et al. Attempted replication of reported chronic obstructive pulmonary disease candidate gene associations. *Am J Respir Cell Mol Biol* 2005;33(1):71-8.
7. O'Connor GT, Sparrow D, Weiss ST. A prospective longitudinal study of methacholine airway responsiveness as a predictor of pulmonary-function decline: the Normative Aging Study. *Am J Respir Crit Care Med* 1995;152(1):87-92.
8. Zhu G, Warren L, Aponte J, Gulsvik A, Bakke P, Anderson WH, et al. The SERPINE2 gene is associated with chronic obstructive pulmonary disease in two large populations. *Am J Respir Crit Care Med* 2007;176(2):167-73.
9. Sood A, Stidley CA, Picchi MA, Celedon JC, Gilliland F, Crowell RE, et al. Difference in airflow obstruction between Hispanic and non-Hispanic White female smokers. *COPD* 2008;5(5):274-81.
10. Pellegrino R, Viegi G, Brusasco V, Crapo RO, Burgos F, Casaburi R, et al. Interpretative strategies for lung function tests. *Eur Respir J* 2005;26(5):948-68.
11. Cho MH, Boutaoui N, Klanderman BJ, Sylvia JS, Ziniti JP, Hersh CP, et al. Variants in FAM13A are associated with chronic obstructive pulmonary disease. *Nat Genet* 2010;42(3):200-2.
12. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *The American Journal of Human Genetics* 2007;81(3):559-75.
13. Abecasis GR, Cherny SS, Cookson WO, Cardon LR. GRR: graphical representation of relationship errors. *Bioinformatics* 2001;17(8):742-3.
14. Miyazawa H, Kato M, Awata T, Kohda M, Iwasa H, Koyama N, et al. Homozygosity Haplotype Allows a Genomewide Search for the Autosomal Segments Shared among Patients. *The American Journal of Human Genetics* 2007;80(6):1090-102.

Supplementary Table 1 – Genotyping quality control for Costa Rica

Illumina Platform	Quad 610
Genotyped Subjects	115
Missingness > 5%	0
Discordance	0
Relatedness	0
Gender mismatch	0
Passing subjects	115
Genotyped markers	619,372
Minor allele frequency = 0 / Missingness > 5%	40,640
Less than 2 subjects with minor allele	16,665
Non-Reproducible	0
HWE < 0.01	3,138
Passing markers	558,929

Supplementary Table 2 – Characteristics of cohorts

	Costa Rica		NETT/NAS,ECLIPSE,Norway combined cohort	
	Cases	Controls	Cases	Controls
No. subjects	58	57	2940	1380
Sex (# male)	33	37	1903	910
Mean age (in years)	60	46	65	60
Mean pack-years	55	35	47	28
Mean FEV1 % predicted (in liters)	43	101	46	98
Mean FEV1/FVC	0.53	0.83	.45	.79

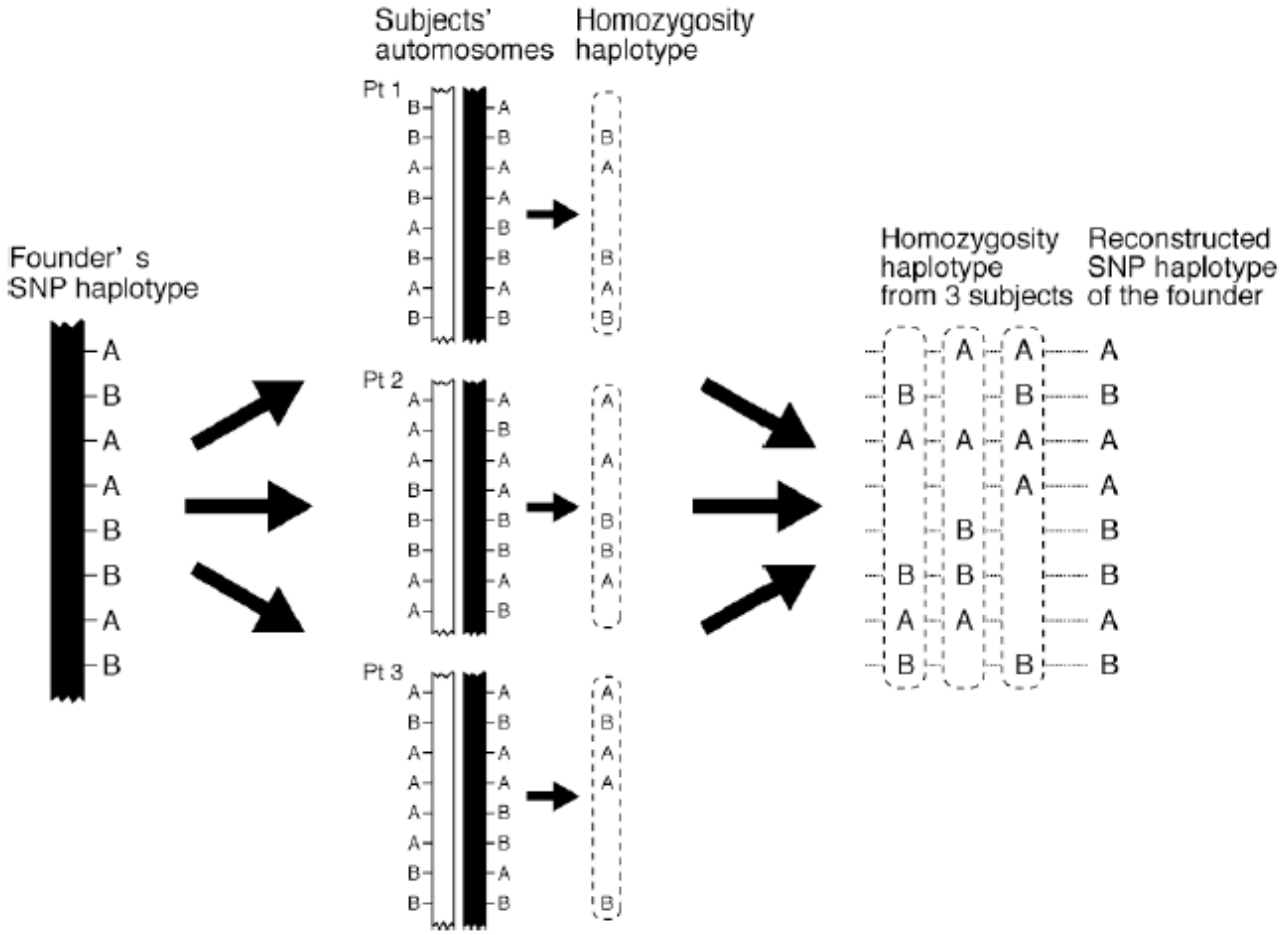
Supplementary Table 3 – Top 20 P values representing 100 RCHH

Location (chromosome and position in base pairs using hg18 coordinates)	Cases N (%)	Controls N (%)	P value	Size (in bp)
6:121788169-122413818	12 (21%)	0 (0%)	2.20E-04	6.3E+05
6:123119324-123274804	12 (21%)	0 (0%)	2.20E-04	1.6E+05
6:123668845-123668845	12 (21%)	0 (0%)	2.20E-04	0.0E+00
7:151892012-151903280	12 (21%)	0 (0%)	2.20E-04	1.1E+04
7:151904404-152295138	14 (24%)	1 (2%)	2.60E-04	3.9E+05
6:121660856-121783240	11 (19%)	0 (0%)	4.20E-04	1.2E+05
6:122414162-123118271	11 (19%)	0 (0%)	4.20E-04	7.0E+05
6:123277300-123664764	11 (19%)	0 (0%)	4.20E-04	3.9E+05
7:151823054-151887401	11 (19%)	0 (0%)	4.20E-04	6.4E+04
15:30302218-30713368	13 (22%)	1 (2%)	5.00E-04	4.1E+05
6:120847633-121657937	10 (17%)	0 (0%)	7.80E-04	8.1E+05
6:123669986-123911605	10 (17%)	0 (0%)	7.80E-04	2.4E+05
7:151524608-151820728	10 (17%)	0 (0%)	7.80E-04	3.0E+05
8:8469618-8476478	10 (17%)	0 (0%)	7.80E-04	6.9E+03
17:4832680-5666242	10 (17%)	0 (0%)	7.80E-04	8.3E+05
15:28124439-29225595	12 (21%)	1 (2%)	9.40E-04	1.1E+06
15:30300525-30301633	12 (21%)	1 (2%)	9.40E-04	1.1E+03
15:30714768-30721385	12 (21%)	1 (2%)	9.40E-04	6.6E+03
6:120356562-120847245	9 (16%)	0 (0%)	1.40E-03	4.9E+05
6:123915133-123935587	9 (16%)	0 (0%)	1.40E-03	2.0E+04
8:8482301-8491941	9 (16%)	0 (0%)	1.40E-03	9.6E+03
15:30748430-30764393	9 (16%)	0 (0%)	1.40E-03	1.6E+04
17:4746831-4830503	9 (16%)	0 (0%)	1.40E-03	8.4E+04
17:5666418-5673154	9 (16%)	0 (0%)	1.40E-03	6.7E+03
20:7112725-7318428	9 (16%)	0 (0%)	1.40E-03	2.1E+05
7:152299155-152322133	11 (19%)	1 (2%)	1.70E-03	2.3E+04
15:28115352-28116500	11 (19%)	1 (2%)	1.70E-03	1.1E+03
15:30723787-30746003	11 (19%)	1 (2%)	1.70E-03	2.2E+04
10:2951101-2969537	16 (28%)	4 (7%)	2.30E-03	1.8E+04
1:54240754-54358003	8 (14%)	0 (0%)	2.70E-03	1.2E+05
6:120203694-120354567	8 (14%)	0 (0%)	2.70E-03	1.5E+05
6:123938996-123954610	8 (14%)	0 (0%)	2.70E-03	1.6E+04
7:151325371-151523278	8 (14%)	0 (0%)	2.70E-03	2.0E+05
8:8493973-8538640	8 (14%)	0 (0%)	2.70E-03	4.5E+04
13:37187973-37868825	8 (14%)	0 (0%)	2.70E-03	6.8E+05
15:30766905-30782048	8 (14%)	0 (0%)	2.70E-03	1.5E+04
20:6808895-7112218	8 (14%)	0 (0%)	2.70E-03	3.0E+05
20:7319819-7499917	8 (14%)	0 (0%)	2.70E-03	1.8E+05
15:29227096-30300468	12 (21%)	2 (4%)	3.10E-03	1.1E+06
7:152326396-152436044	10 (17%)	1 (2%)	3.20E-03	1.1E+05
8:8450177-8467882	10 (17%)	1 (2%)	3.20E-03	1.8E+04
15:28091067-28113480	10 (17%)	1 (2%)	3.20E-03	2.2E+04
21:16580023-16650095	10 (17%)	1 (2%)	3.20E-03	7.0E+04
10:2969826-2976450	15 (26%)	4 (7%)	4.00E-03	6.6E+03

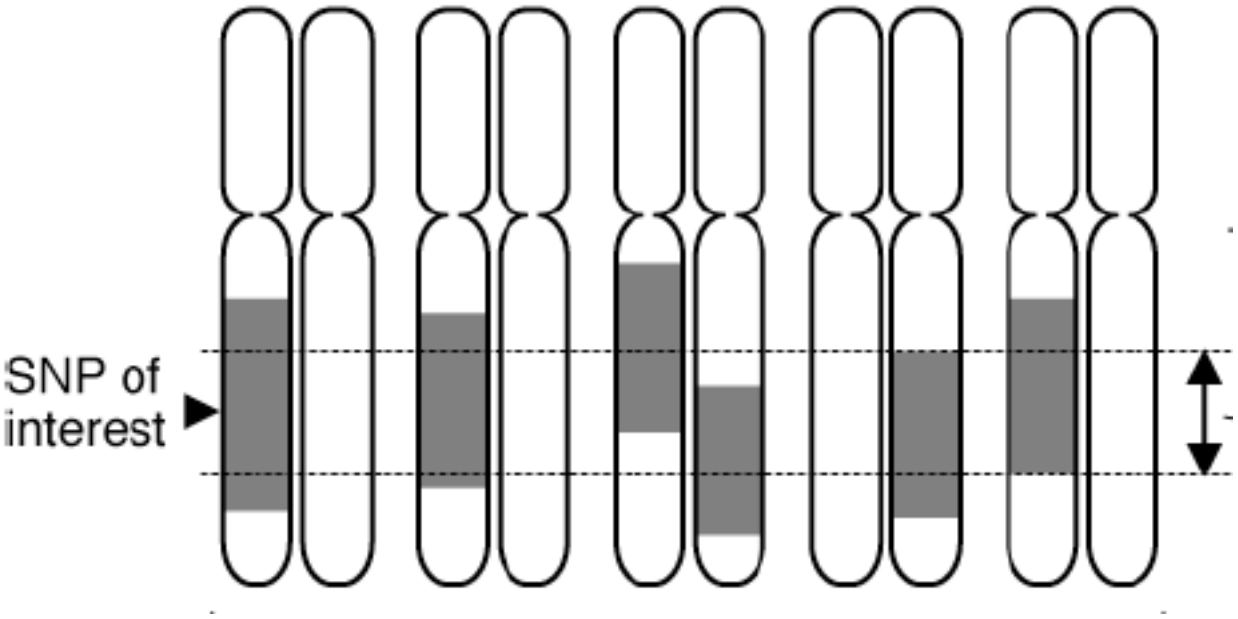
8:8405647-8411849	13 (22%)	3 (5%)	4.90E-03	6.2E+03
1:54359237-54379392	7 (12%)	0 (0%)	4.90E-03	2.0E+04
3:195215725-195237176	7 (12%)	0 (0%)	4.90E-03	2.1E+04
6:20108027-21710531	7 (12%)	0 (0%)	4.90E-03	1.6E+06
6:120164641-120203646	7 (12%)	0 (0%)	4.90E-03	3.9E+04
6:158414763-159470841	7 (12%)	0 (0%)	4.90E-03	1.1E+06
6:166716423-166938711	7 (12%)	0 (0%)	4.90E-03	2.2E+05
7:106220834-109629284	7 (12%)	0 (0%)	4.90E-03	3.4E+06
7:130062836-130088545	7 (12%)	0 (0%)	4.90E-03	2.6E+04
7:151325334-151325334	7 (12%)	0 (0%)	4.90E-03	0.0E+00
8:8538732-8577480	7 (12%)	0 (0%)	4.90E-03	3.9E+04
8:141202813-142094332	7 (12%)	0 (0%)	4.90E-03	8.9E+05
11:58979902-61009776	7 (12%)	0 (0%)	4.90E-03	2.0E+06
11:125802819-126097478	7 (12%)	0 (0%)	4.90E-03	2.9E+05
11:126163084-126355974	7 (12%)	0 (0%)	4.90E-03	1.9E+05
13:21298032-21305888	7 (12%)	0 (0%)	4.90E-03	7.9E+03
13:36237281-36283266	7 (12%)	0 (0%)	4.90E-03	4.6E+04
13:37140481-37185893	7 (12%)	0 (0%)	4.90E-03	4.5E+04
13:37873623-37973865	7 (12%)	0 (0%)	4.90E-03	1.0E+05
14:20026284-20062727	7 (12%)	0 (0%)	4.90E-03	3.6E+04
14:20117472-20117472	7 (12%)	0 (0%)	4.90E-03	0.0E+00
15:30782135-30810778	7 (12%)	0 (0%)	4.90E-03	2.9E+04
15:45975580-49182801	7 (12%)	0 (0%)	4.90E-03	3.2E+06
15:56434430-56450814	7 (12%)	0 (0%)	4.90E-03	1.6E+04
17:1738924-1758733	7 (12%)	0 (0%)	4.90E-03	2.0E+04
17:4636312-4742067	7 (12%)	0 (0%)	4.90E-03	1.1E+05
18:6634394-6680637	7 (12%)	0 (0%)	4.90E-03	4.6E+04
18:6830148-6841557	7 (12%)	0 (0%)	4.90E-03	1.1E+04
18:70392209-70397061	7 (12%)	0 (0%)	4.90E-03	4.9E+03
18:70645532-70974303	7 (12%)	0 (0%)	4.90E-03	3.3E+05
18:72019892-72356083	7 (12%)	0 (0%)	4.90E-03	3.4E+05
19:56106962-56106962	7 (12%)	0 (0%)	4.90E-03	0.0E+00
20:7499995-7572747	7 (12%)	0 (0%)	4.90E-03	7.3E+04
10:2947286-2950317	16 (28%)	5 (9%)	5.40E-03	3.0E+03
7:152436388-152475034	9 (16%)	1 (2%)	5.80E-03	3.9E+04
13:21279294-21279294	9 (16%)	1 (2%)	5.80E-03	0.0E+00
15:28085029-28088527	9 (16%)	1 (2%)	5.80E-03	3.5E+03
17:5674078-5758986	9 (16%)	1 (2%)	5.80E-03	8.5E+04
10:2977064-2978202	14 (24%)	4 (7%)	6.90E-03	1.1E+03
8:8412246-8427477	12 (21%)	3 (5%)	8.50E-03	1.5E+04
10:2719235-2728665	12 (21%)	3 (5%)	8.50E-03	9.4E+03
13:21234739-21239466	12 (21%)	3 (5%)	8.50E-03	4.7E+03
21:46252267-46265743	12 (21%)	3 (5%)	8.50E-03	1.3E+04
1:54379745-54491077	6 (10%)	0 (0%)	9.00E-03	1.1E+05
1:145750325-149855359	6 (10%)	0 (0%)	9.00E-03	4.1E+06
1:163312158-163357909	6 (10%)	0 (0%)	9.00E-03	4.6E+04
2:80955775-83248796	6 (10%)	0 (0%)	9.00E-03	2.3E+06
2:221397282-221749310	6 (10%)	0 (0%)	9.00E-03	3.5E+05
3:195179804-195213553	6 (10%)	0 (0%)	9.00E-03	3.4E+04

4:6644018-6644018	6 (10%)	0 (0%)	9.00E-03	0.0E+00
4:6818237-7317757	6 (10%)	0 (0%)	9.00E-03	5.0E+05
4:7381577-7389520	6 (10%)	0 (0%)	9.00E-03	7.9E+03
4:7531322-7990094	6 (10%)	0 (0%)	9.00E-03	4.6E+05
6:19963912-20104787	6 (10%)	0 (0%)	9.00E-03	1.4E+05
6:21710724-23614469	6 (10%)	0 (0%)	9.00E-03	1.9E+06
6:51100102-53152338	6 (10%)	0 (0%)	9.00E-03	2.1E+06

Supplementary figure 1 – The homozygosity haplotype of a theoretical founder individual is reconstructed using the algorithm shown here. While the three genotyped subjects themselves are not homozygous the region shown here, by comparing across all subjects the haplotype of the founder can be reconstructed.



Supplementary figure 2 – The genetic length of a shared segment between any two individuals must be at least 3.0 cM, but the total length of the final RCHH may be much smaller due to differences in overlap. In this figure, the region derived from a common ancestor is shaded grey. To be identified by the algorithm, any two subjects must have an overlap of common ancestry that is at least 3.0 cM in length. However, the final reconstructed Region from a Common Ancestor using all subjects (indicated as the region between the two dashed lines) may be much smaller than 3.0 cM.



Supplementary figure 3 – In the case where more than one region from a common ancestor is identified, the region with the largest number of subjects in the subgroup is used. In the figure below, 5 subjects inherited the same region on at least one chromosome (dark grey), and two subjects inherited a different region (light grey). The HHAnalysis algorithm chooses the dark grey region as the ancestral segment.