ORIGINAL ARTICLE

Genetic regulation of gene expression in the lung identifies *CST3* and *CD22* as potential causal genes for airflow obstruction

Maxime Lamontagne,[1] Wim Timens,[2] Ke Hao,[3] Yohan Bossé,[1,4] Michel Laviolette,[1] Katrina Steiling,[5] Joshua D Campbell,[5] Christian Couture,[1] Massimo Conti,[1] Karen Sherwood,[6] James C Hogg,[6,7] Corry-Anke Brandsma,[2] Maarten van den Berge,[8] Andrew Sandford,[6,9] Stephen Lam,[10] Marc E Lenburg,[5] Avrum Spira,[5] Peter D Paré,[6,9] David Nickle,[11] Don D Sin,[6,9] Dirkje S Postma[8]

▶ Additional material is published online only. To view please visit the journal online (http://dx.doi.org/10.1136/thoraxjnl-2014-205630).

For numbered affiliations see end of article.

Correspondence to
Professor Dirkje S Postma, University of Groningen, University Medical Center Groningen, Department of Pulmonary Medicine and Tuberculosis, GRIAC Research Institute, AA11, Hanzeplein, PO Box 30001, Groningen 9700 RB, The Netherlands; d.s.postma@umcg.nl

ML, WT, and KH contributed equally.

Received 28 April 2014
Revised 16 July 2014
Accepted 14 August 2014
Published Online First
2 September 2014

ABSTRACT

Background COPD is a complex chronic disease with poorly understood pathogenesis. Integrative genomic approaches have the potential to elucidate the biological networks underlying COPD and lung function. We recently combined genome-wide genotyping and gene expression in 1111 human lung specimens to map expression quantitative trait loci (eQTL).

Objective To determine causal associations between COPD and lung function-associated single nucleotide polymorphisms (SNPs) and lung tissue gene expression changes in our lung eQTL dataset.

Methods We evaluated causality between SNPs and gene expression for three COPD phenotypes: FEV_1% predicted, FEV_1/FVC and COPD as a categorical variable. Different models were assessed in the three cohorts independently and in a meta-analysis. SNPs associated with a COPD phenotype and gene expression were subjected to causal pathway modelling and manual curation. *In silico* analyses evaluated functional enrichment of biological pathways among newly identified causal genes. Biologically relevant causal genes were validated in two separate gene expression datasets of lung tissues and bronchial airway brushings.

Results High reliability causal relations were found in SNP–mRNA–phenotype triplets for FEV_1% predicted (n=169) and FEV_1/FVC (n=80). Several genes of potential biological relevance for COPD were revealed. eQTL-SNPs upregulating cystatin C (*CST3*) and *CD22* were associated with worse lung function. Signalling pathways enriched with causal genes included xenobiotic metabolism, apoptosis, protease–antiprotease and oxidant–antioxidant balance.

Conclusions By using integrative genomics and analysing the relationships of COPD phenotypes with SNPs and gene expression in lung tissue, we identified *CST3* and *CD22* as potential causal genes for airflow obstruction. This study also augmented the understanding of previously described COPD pathways.

Key messages

What is the key question?

- What are the causal genetic variants changing gene expression in the lung that in turn associate with lower lung function and COPD?

What is the bottom line?

- Lung function-associated genetic variants alter the mRNA expression of nearby genes involved in biological pathways underpinning pulmonary function and COPD pathogenesis.

Why read on?

- New genes of airflow obstruction are identified with a generalised framework for the identification of causal genes from joint examination of genome-wide genotyping and gene expression data in the same patients.

INTRODUCTION

Genome-wide association studies (GWAS) have revolutionised our ability to identify common genetic variants that are associated with complex chronic diseases.[1] This approach has been applied to COPD, a lung disease that is caused predominantly by cigarette smoking in the western world.[2–4] It is well known that only a subset of heavy smokers (15–20%) develop clinically relevant COPD and there is considerable evidence that there is a substantial genetic component involved in its pathogenesis.[5] Although GWAS have identified novel loci that harbour susceptibility genes, they do not allow precise identification of the causal variant (or variants). In addition, GWAS do not provide information on how and to what extent the gene (or genes) within the susceptibility loci contribute to the phenotype. Interestingly, the majority of genetic variants which have been associated with disease traits by GWAS do not affect the coding sequence of genes but are located in intergenic regions or introns.[6] Possible explanations are that the associated alleles are in linkage disequilibrium (LD) with rarer coding alleles with large effect sizes[7] and/or that the genetic variants control the level of expression of genes involved in pathogenetic pathways. For complex genetic diseases, such as COPD, the effects of susceptibility alleles may primarily act by regulating gene

To cite: Lamontagne M, Timens W, Hao K, *et al*. *Thorax* 2014;**69**:997–1004.

expression rather than by altering protein coding as in most Mendelian diseases.[8]

We recently reported the discovery of a large number of lung-specific expression quantitative trait loci (eQTLs)[9] and identified the most likely causal genes within three GWAS-nominated COPD susceptibility loci.[10] The aim of the present study is to use the power of genome-wide mRNA expression arrays combined with genome-wide interrogation of single nucleotide polymorphisms (SNPs) to pinpoint specific SNPs that are related to lung tissue gene expression and to COPD phenotypes. Identification of susceptibility alleles that function as strong eQTLs increases the likelihood of identifying the true susceptibility gene within loci with large areas of LD.[8] Moreover, the use of integrative genomics by combining environmental exposure data with susceptibility alleles, RNA expression levels, and different COPD phenotypes can unravel causal genetic relationships.[11] The basic principle of the current study is to map the genetic regulation of gene expression to identify DNA variants that induce changes in transcriptional networks that in turn contribute to COPD and airway obstruction pathogenesis.

METHODS

Subject selection

The methods for subject selection and phenotyping, and for interrogation of gene expression and genotype were recently described.[9] The lung tissue used for discovery of eQTLs was from 1111 human subjects who underwent lung surgery at three academic sites, Laval University, University of British Columbia (UBC) and University of Groningen, henceforth referred to as Laval, UBC and Groningen, respectively. All lung specimens from Laval were obtained from patients undergoing lung cancer surgery and were harvested from a site distant from the tumour. At UBC, the majority of samples were from patients undergoing resection of small peripheral lung lesions. Additional samples were from autopsy and at the time of lung transplantation. At Groningen, the lung specimens were obtained at surgery from patients with various lung diseases, including patients undergoing therapeutic resection for lung tumours, harvested from a site distant from the tumour, and lung transplantation. For the present study, the principal aim was to examine smoking-related airway obstruction. Thus, we excluded subjects whose lung function may have been influenced by lung diseases other than COPD and lung cancer. Exclusion criteria are provided in the online supplement.

COPD phenotypes and GWAS

Genome-wide association was performed using linear or logistic regression models on the three phenotypes: FEV_1% predicted and FEV_1/FVC as continuous variables, and COPD defined dichotomously based on an $FEV_1/FVC<0.7$ cut-off (see online supplement). Single-marker association tests were run within each cohort adjusting for age, gender and smoking status. Fixed-effects meta-analysis was then performed combining the three cohorts using inverse SE weighting.

Expression trait processing

Expression traits were adjusted for age, gender and smoking status as described previously.[9] Gene expression data are available in the Gene Expression Omnibus repository through accession number GSE23546.

Causality models

We evaluated three competing causality models to describe the relationships between lung eQTL-SNPs, RNA expression and COPD phenotypes[12] (figure 1). Model 1 indicates a potential causal relationship where the SNP acts on gene expression to produce the phenotype, which was the main interest of this work. Model 2 indicates that the gene expression pattern is reactive to the phenotype, that is, the phenotype drives gene expression. Model 3 is the independent model where the SNP acts on the phenotype and gene expression independently. We required a p value $<1\times10^{-3}$ for SNP associated with phenotype before examining a particular variable triplet (ie, SNP, gene expression and phenotypes). The causal model (model 1) was selected when p value <0.05 was found for SNP associated with gene expression adjusting for phenotype and p value >0.05 was found for SNP associated with phenotype adjusting for gene expression. More details are provided in the online supplement. Analyses were performed in the three cohorts separately and then combined into a meta-analysis. We then conducted sample bootstrapping and repeated the causality test for N=1000 realisations. The reliability score of the molecular relationship is the fraction of bootstrap realisations that supports the call observed. Threshold for p values, number of bootstrap reselections and reliability cut-off (>0.8) were selected based on previous literature.[12]

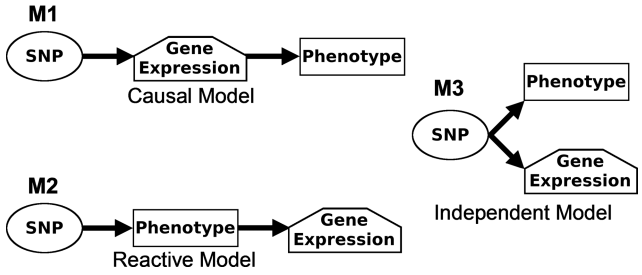


Figure 1 Causality models showing the relationships between the expression of a gene, a phenotype and a single nucleotide polymorphism (SNP). Three models are depicted including a causal model (M1), a reactive model (M2), and an independent model (M3). M1 is the simplest, and states that the genotype at an expression quantitative trait loci SNP acts directly on gene expression pattern to produce the phenotype. M2 states that the gene expression pattern is reactive to the phenotype, and M3 states that the gene expression pattern and phenotype are independent.

Manual curation and pathway analyses on reliable causal models

The biology and possible role of genes in the causal model were reviewed by manual curation and bioinformatics tools, including Ingenuity Pathway Analysis (IPA), MetaCore and Partek (see online data supplement).

Replication studies

Biologically relevant causal genes were validated in two datasets. First, a bronchial airway epithelial dataset where genome-wide gene expression levels were obtained from bronchial brushing of 238 individuals and associated with lung function as previously described.[13] Second, a regional lung tissue dataset where genome-wide gene expression was obtained for eight regions of the same lung and for eight patients. The association between gene expression levels and micro-CT-based regional emphysema severity was assessed.[14] More details are provided in the online data supplement.

RESULTS

Of the 1111 subjects in whom there were data on genotype and gene expression, 848 with sufficient phenotypic information

Table 1 Clinical characteristics of patients

Characteristics	Laval (n=403)	UBC (n=270)	Groningen (n=175)
Age (years)	63.4±9.8	63.9±10.0	59.7±10.0
Male/female (n)	224/179	144/126	83/92
Body mass index (kg/m^2)	26.6±5.2	25.7±5.4	24.6±4.1
FEV_1% predicted	82.2±17.6	83.6±22.3	72.4±24.9
FEV_1/FVC	0.68±0.10	0.69±0.13	0.64±0.16
COPD (%)	209 (51.9)	114 (42.2)	112 (64.0)
Stage 1: mild	80 (38.2)	43 (37.7)	16 (14.3)
Stage 2: moderate	117 (56.0)	60 (52.6)	30 (26.8)
Stage 3: severe	11 (5.3)	2 (1.8)	11 (9.8)
Stage 4: very severe	1 (0.5)	9 (7.9)	44 (39.3)
Missing data	0	0	11 (9.8)
Non-COPD (%)	194 (48.1)	156 (57.8)	63 (36.0)
Smoking (%)			
Smoker	89 (22.1)	91 (33.7)	12 (6.9)
Ex-smoker	281 (69.7)	150 (55.6)	117 (66.9)
Non-smoker	33 (8.2)	16 (5.9)	43 (24.6)
Missing data	0	13 (4.8)	3 (1.7)
Pack years (years)	48.6±27.4	46.0±28.6	32.1±18.1

Continuous variables are presented as means±SDs.
UBC, University of British Columbia.

were included in the analysis. The demographic and clinical features of the subjects in the three cohorts are described in table 1.

Causal pathways that fit model 1

Model fitting was restricted to SNPs that were significantly associated with at least one of the lung function variables or COPD as a categorical variable at a p value cut-off of 10^{-3}. SNPs were considered an eQTL if they had a p value $\leq 10^{-5}$. Using these criteria, there were 4465 SNP–mRNA–phenotype triplets. Of these, 249 triplets showed a significant fit to the causal model for at least one of the phenotypes with a reliability score >0.8 in at least one cohort and/or in the meta-analysis. A fit to this model means that the SNP was associated with one of the phenotypes and affected gene expression in a direction that supported a causal relationship.

The list of causal models with a confidence score >0.8 in any one of the cohorts or in the meta-analysis is shown in online supplementary table S1. For FEV_1% predicted, 169 causal models were found, including 122 unique SNPs and 169 probe sets. The 169 models included 168 *cis* eQTLs and 1 *trans* eQTL, respectively. For FEV_1/FVC, 80 causal models were found involving 63 unique SNPs and 80 probe sets. Among these 80 models, 79 were *cis* eQTLs and 1 was a *trans* eQTL. No causal model was found when using COPD as a categorical variable. There was no overlap between causal pathway models for FEV_1% predicted and FEV_1/FVC.

Manual curation of significant and reliable causal models

Significant causal models were inspected manually. We identified a number of models of potential biological relevance for FEV_1% predicted and FEV_1/FVC. The most biologically relevant models are listed in table 2. The p values and direction of effect for each of the SNP associations with the gene expression and the phenotype are also indicated. For example, SNP rs6048956 was significantly associated with cystatin C (*CST3*) transcript. The common allele was associated with higher expression of the transcript (figure 2 and positive eQTL Z score in table 2) and with lower FEV_1% predicted (figure 2 and negative Z score with phenotype in table 2). Taken together, these results suggest that the common allele confers susceptibility to a lower FEV_1% predicted value through upregulation of the *CST3* mRNA expression levels in the lung. This was confirmed in a second causality model that interrogated a different probe

Table 2 Models of biological relevance identified by manual curation

SNPs	Reference allele (freq)*	Gene (probe set)	Lung eQTL p value†	eQTL Z score‡	p Value phenotype§	Z score phenotype¶
FEV_1% predicted						
rs769178	G (0.91)	*NCR3* (100125842_TGI_at)	8.03×10^{-30}	11.3	4.7×10^{-4}	3.5
rs6048956	C (0.77–0.79)	*CST3* (100307577_TGI_at)	1.54×10^{-68}	17.5	7.8×10^{-4}	−3.4
rs6515375	G (0.79–0.81)	*CST3* (100125967_TGI_at)	3.65×10^{-6}	4.6	5.8×10^{-4}	−3.4
rs2270859	G (0.83–0.88)	*CSTA* (100148334_TGI_at)	1.74×10^{-9}	6.0	3.1×10^{-6}	−4.7
rs4550905	G (0.26–0.31)	*PPARGC1A* (100131093_TGI_at)	1.69×10^{-5}	−4.3	4.4×10^{-4}	−3.5
rs3803761	G (0.71–0.77)	*FLCN* (100135396_TGI_at)	3.77×10^{-29}	11.2	3.4×10^{-2}	−2.1
rs1543438	A (0.76–0.82)	*BCL2L1* (100158784_TGI_at)	3.89×10^{-5}	−4.1	7.94×10^{-5}	3.9
rs9880397	G (0.62–0.64)	*CADM2* (100162763_TGI_at)	1.3×10^{-8}	−5.7	2.9×10^{-4}	3.6
rs2466183	T (0.84–0.86)	TNFRSF10B (100153254_TGI_at)	3.76×10^{-5}	−4.1	6.6×10^{-4}	3.4
FEV_1/FVC						
rs17754977	A (0.30–0.32)	*GSTO2* (100132911_TGI_at)	3.96×10^{-5}	4.1	8.2×10^{-4}	3.3
rs9987135	T (0.28–0.31)	*DEPDC6* (100154484_TGI_at)	6.13×10^{-59}	16.2	1.1×10^{-4}	−3.9
rs10411704	T (0.79–0.82)	*CD22* (100154732_TGI_at)	1.70×10^{-40}	−13.3	4.6×10^{-4}	3.5
rs12179536	A (0.80–0.84)	*MUC22* (100304000_TGI_at)	3.47×10^{-33}	−12.0	3.0×10^{-3}	3.0
rs2287765	T (0.91–0.93)	*SPINK5* (100305138_TGI_at)	2.85×10^{-10}	6.3	9.2×10^{-4}	−3.3

p Values and Z scores in this table are from the meta-analysis. As indicated in the text, the significant causal models were selected based on results of both individual cohorts and meta-analysis (SNP associated with phenotype with p value $<10^{-3}$, SNPs associated with gene expression with p value $\leq10^{-5}$, and reliability score >0.8 in at least one cohort and/or in the meta-analysis). Known role of genes are provided in online supplementary table S2.
*Frequency of the reference allele in the three cohorts.
†Lung eQTL p-value from the meta-analysis.
‡Z score from the eQTL meta-analysis showing the direction of effect for the SNP on gene expression.
§p Value for association between the SNP and phenotype from the meta-analysis.
¶Effect size (Z score) for association between the SNP and the phenotype from the meta-analysis, showing the direction of effect for the SNP on the phenotype.
eQTL, expression quantitative trait loci; SNP, single nucleotide polymorphism.

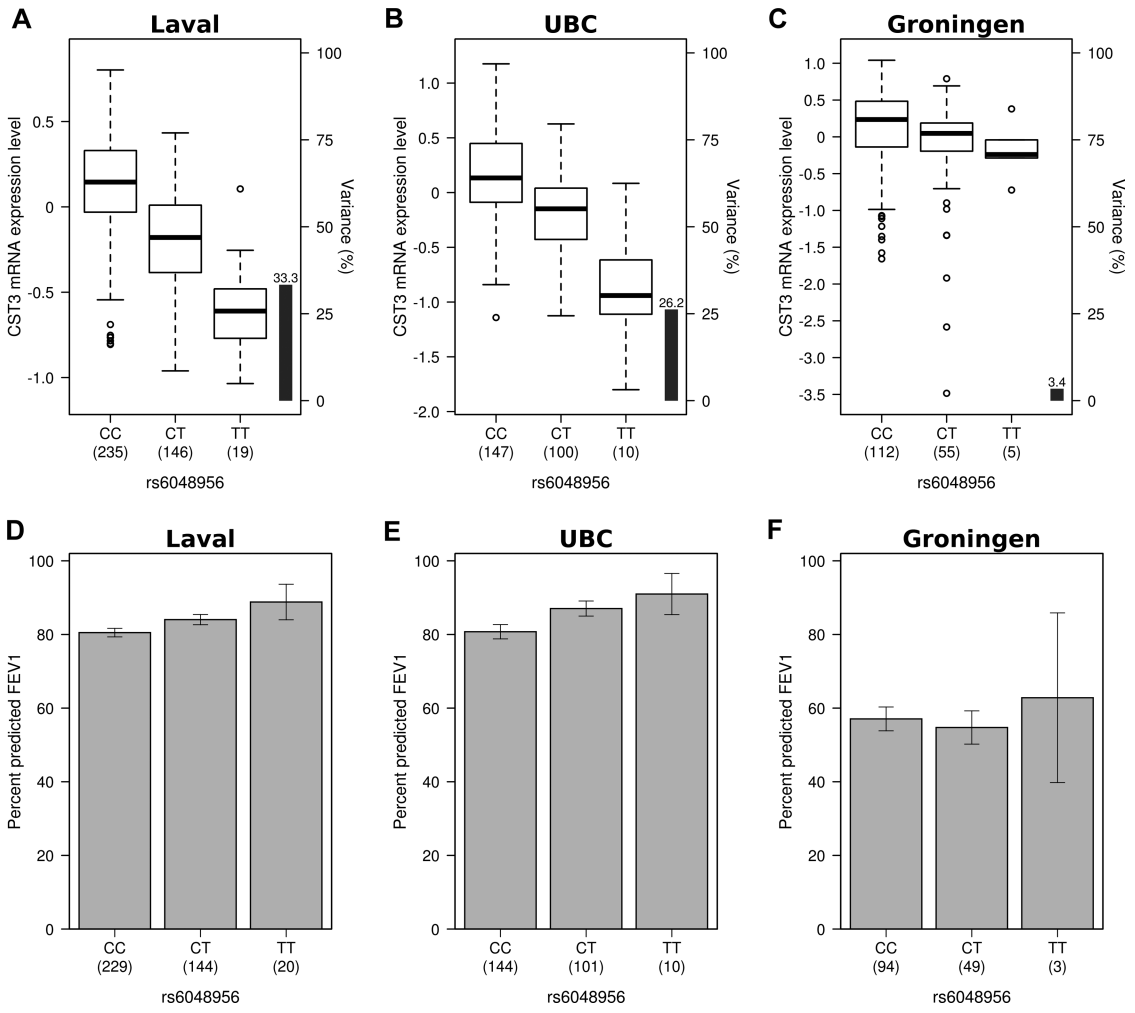


Figure 2 Direction of effects for causality model with rs6048956, mRNA expression of cystatin C (*CST3*) and FEV_1% predicted. The left, centre and right panels show the results for Laval, University of British Columbia (UBC) and Groningen samples, respectively. (A–C) are boxplots of gene expression levels in the lung for *CST3* according to genotype groups for single nucleotide polymorphism (SNP) rs6048956. The left y-axis shows the mRNA expression levels for *CST3*. The x-axis represents the three genotype groups for SNP rs6048956. The right y-axis shows the proportion of the gene expression variance explained by the SNP (black bar). Box boundaries, whiskers and centre mark in boxplots represent the first and third quartiles, the most extreme data point which is no more than 1.5 times the IQR, and median, respectively. (D–F) are barplots of the mean and SE of FEV_1% predicted according to genotype groups for SNP rs6048956.

set for *CST3*, FEV_1% predicted and rs6515375 (table 2 and see online supplementary figure S1). Analogous observations were made for natural cytotoxicity triggering receptor 3 (*NCR3*), cystatin A (*CSTA*), peroxisome proliferation-activated receptor γ coactivator 1α (*PPARGC1A*), folliculin (*FLCN*), BCL2-like 1 (*BCL2L1*), cell adhesion molecule 2 (*CADM2*), and tumour necrosis factor receptor superfamily member 10b (*TNFRSF10B*) for FEV_1% predicted, and glutathione S-transferase ω 2 (*GSTO2*), DEP domain containing 6 (*DEPDC6*), CD22 molecule (*CD22*), mucin 22 (*MUC22*), and serine peptidase inhibitor Kazal type 5 (*SPINK5*) for FEV_1/FVC (table 2). The direction of effects for all these causality models are illustrated in online supplementary figures S2–13.

Pathway analyses

To find key biological pathways involved in COPD, causality genes were overlaid on canonical pathways available in IPA. One hundred and sixty-nine genes and 80 causality genes for FEV_1 and FEV_1/FVC were considered, respectively (see online supplementary table S1). Table 3 shows all canonical pathways enriched with causality genes ($p<0.05$). Of interest was the aryl hydrocarbon receptor signalling pathway, which is involved in xenobiotic clearance. Five causality genes were noted in this canonical pathway, including *ARNT* (aryl hydrocarbon receptor nuclear translocator), *IL6* (interleukin 6), *GSTO2*, *ALDH8A1* (aldehyde dehydrogenase 8 family, member A1) and *TRIP11* (thyroid hormone receptor interactor 11). The aryl hydrocarbon receptor signalling pathway and the location of the causality genes are illustrated in online supplementary figure S14. Another pathway involved in xenobiotic handling, the xenobiotic metabolism signalling pathway, was also enriched for causality genes, many of which overlapped with those in the aryl hydrocarbon receptor signalling pathway (*ARNT*, *IL6*, *GSTO2* and *ALDH8A1*). The former also includes *PPARGC1A* and *PRKCE* (protein kinase C, ε). This pathway is illustrated in online supplementary figure S15.

For the Partek analyses 118 and 45 transcripts were identified for FEV_1% predicted and FEV_1/FVC, respectively. These transcripts mapped to a total of 115 genes in Partek GS (based on official gene symbols) and were parsed to Partek Pathway Suite. Causality genes were overlaid onto canonical pathways from the REACTOME and KEGG databases. Table 4 shows the canonical

Table 3 Canonical pathways enriched for causality genes using the Ingenuity Pathway Analysis system

Pathways	p Value
Cdc42 signalling	0.00002
Crosstalk between dendritic cells and natural killer cells	0.00015
Graft-versus-host disease signalling	0.00069
OX40 signalling pathway	0.00151
Antigen presentation pathway	0.00468
Neuroprotective role of THOP1 in Alzheimer's disease	0.00589
Taurine and hypotaurine metabolism	0.00617
Autoimmune thyroid disease signalling	0.00661
Regulation of actin-based motility by Rho	0.00676
Communication between innate and adaptive immune cells	0.00676
Aryl hydrocarbon receptor signalling	0.00759
Integrin signalling	0.00891
Allograft rejection signalling	0.00977
Cytotoxic T-lymphocyte-mediated apoptosis of target cells	0.01023
Cyanoamino acid metabolism	0.01047
Glutathione metabolism	0.01072
Amyotrophic lateral sclerosis signalling	0.01230
Systemic lupus erythematosus signalling	0.01230
PXR/RXR activation	0.01995
Macropinocytosis signalling	0.02455
Selenoamino acid metabolism	0.02692
Xenobiotic metabolism signalling	0.02951
Cellular effects of sildenafil (Viagra)	0.03162
Aminoacyl-tRNA biosynthesis	0.03631
Actin cytoskeleton signalling	0.04365
FAK signalling	0.04467
Apoptosis signalling	0.04677

Pathways discussed in the text are given in italic.

Table 4 Canonical pathways enriched for causality genes using Partek

Pathways	p Value	Pathway ID
TRAIL signalling	0.0003	reactome_pathway_506
Cyanoamino acid metabolism	0.001	kegg_pathway_36
Cell adhesion molecules	0.002	kegg_pathway_139
Extrinsic pathway for apoptosis	0.002	reactome_pathway_502
Death receptor signalling	0.002	reactome_pathway_503
Taurine and hypotaurine metabolism	0.002	kegg_pathway_34
Glutathione metabolism	0.003	kegg_pathway_39
Aminoacyl-tRNA biosynthesis	0.004	kegg_pathway_81
Apoptosis	0.006	reactome_pathway_501
Cytosolic tRNA aminoacylation	0.006	reactome_pathway_1127
Glutathione conjugation	0.007	reactome_pathway_256
Natural killer cell mediated cytotoxicity	0.011	kegg_pathway_152
The NLRP1 inflammasome	0.015	reactome_pathway_469
tRNA aminoacylation	0.018	reactome_pathway_1126
Apoptosis	0.023	kegg_pathway_126
BoNT light chain types B, D and F cleave VAMP/synaptobrevin	0.024	reactome_pathway_647
Amyloids	0.026	reactome_pathway_643
Nucleotide-binding domain, leucine rich repeat containing receptor (NLR) signalling pathways	0.026	reactome_pathway_465
BH3-only proteins associate with and inactivate anti-apoptotic BCL-2 members	0.034	reactome_pathway_517
Axonal growth inhibition (RHOA activation)	0.039	reactome_pathway_984
Vitamin C (ascorbate) metabolism	0.039	reactome_pathway_188
Import of palmitoyl-CoA into the mitochondrial matrix	0.039	reactome_pathway_70
Downregulation of ERBB4 signalling	0.039	reactome_pathway_1023
p75NTR regulates axonogenesis	0.043	reactome_pathway_982
Caspase-8 is formed from procaspase-8	0.043	reactome_pathway_507
Activation of procaspase-8	0.043	reactome_pathway_508
Endosomal/vacuolar pathway	0.043	reactome_pathway_413
Phase II conjugation	0.047	reactome_pathway_249
SLBP independent processing of histone pre-mRNAs	0.048	reactome_pathway_1111
Arachidonic acid metabolism	0.050	kegg_pathway_56

Pathways in italic are also found in ingenuity pathway analysis.
TRAIL, tumour necrosis factor related apoptosis-inducing ligand.

pathways enriched with causality genes ($p<0.05$). TNF-related apoptosis-inducing ligand (TRAIL) signalling was the most significant pathway, with several other apoptosis-related pathways included in the list of top significant associations. The glutathione metabolism pathway connects to both the cyanoamino acid and taurine and hypotaurine metabolism pathways (see online supplementary figure S16). These three pathways were identified in our enrichment analysis and were also identified using the IPA system. GO functional categories also showed enrichment for causality genes (see online supplementary figure S17) including the γ-glutamyl transferase activity.

Replication of the most biologically relevant causality models

No other large-scale lung eQTL dataset is available in patients with and without COPD. To replicate the most biologically relevant causality models, we relied on two genome-wide expression datasets. Causality genes were first validated in a gene expression study of bronchial airway epithelial cells obtained by bronchoscopy. All 13 genes represented in table 2, except *MUC22*, were assayed in this airway dataset. Two genes were significantly associated with lung function at a false discovery rate (FDR) of 5% in the bronchial airway epithelial dataset: *CSTA* (FDR=0.002 with FEV_1% predicted) and *TNFRSF10B* (FDR=9.35×10^{-5} with FEV_1% predicted). In both cases, the direction of effect was the same as that in the current study (table 5). For example, higher *CSTA* mRNA levels were associated with worse lung function (table 2 and see online supplementary figure S3).

Causality genes identified in this study were also compared with a second lung transcriptomic study that evaluated the impact of regional emphysema severity on gene expression. Interestingly, *CD22* was positively associated with regional emphysema severity within individuals ($p=5.8\times10^{-5}$) and was a part of the 127 gene signature for emphysema identified in that study.[14] This observation is consistent with the current study showing that carriers of the rare allele for rs10411704 had greater mRNA expression of *CD22* and worse FEV_1/FVC (table 2 and see online supplementary figure S11). Associations with regional emphysema severity were also observed for three other genes in table 2 including *NCR3* ($p=0.058$), *PPARGC1A* ($p=0.081$) and *BCL2L1* ($p=0.051$), but these were not statistically significant. However, the direction of effect was consistent only for *PPARGC1A*. The expression of this gene was found to decrease with emphysema severity in the regional lung tissue dataset and, in the current study, carriers of the rare allele for

Table 5 Replication of causality genes in the bronchial airway epithelial and the regional lung tissue datasets

	Replication datasets	
	Bronchial airway epithelium[13]	Regional lung tissue[14]
FEV_1% predicted		
NCR3	True	False
CST3	True	
CSTA	True*	
PPARGC1A	False	True
FLCN	False	
BCL2L1	True	False
CADM2	False	
TNFRSF10B	True*	
FEV_1/FVC		
GSTO2	False	
DEPDC6	False	
CD22	True	True*
MUC22		
SPINK5	True	

*Significant in the original studies.[13 14]

rs4550905 were associated with less mRNA expression of *PPARGC1A* in the lung and lower FEV_1 (table 2 and see online supplementary figure S4).

DISCUSSION

This investigation integrated a genome-wide eQTL study on lung tissue with genome-wide genetic association results for lung function and COPD in the same subjects. For the discovery of eQTLs we used the entire dataset which consisted of 1111 tissue samples from subjects who had lung surgery for a variety of reasons. In order to focus on COPD, we limited the study to 848 subjects with sufficient phenotypic information and without a lung disease (other than COPD and lung cancer) which could cause abnormalities of pulmonary function. We limited the causal pathway analysis to SNPs that were both significantly related to gene expression ($p<10^{-5}$) and were associated with one of three COPD phenotypes ($p<10^{-3}$), that is, FEV_1% predicted, FEV_1/FVC and COPD defined as $FEV_1/FVC<0.7$. Of the 4465 SNP–mRNA–phenotype triplets which met our inclusion criteria, 249 triplets showed a significant fit to the causality model for at least one of the phenotypes with a reliability score ≥0.8. We thus provide evidence that these SNPs influence disease susceptibility by altering gene expression. Causality pathway genes were enriched in pathways involved in xenobiotic handling, antiprotease and antioxidant activity and apoptosis.

Two of the eQTL-SNPs in *CST3* (rs6515375 and rs6048956) and another for *CSTA* (rs2270859) were in the causal pathway for FEV_1% predicted. The two *CST3* SNPs were in perfect LD and were eQTLs for two different probe sets. *CST3* and *CSTA* are cysteine antiproteases; *CST3* antagonises cysteine cathepsins such as cathepsin L and S, and *CSTA* acts similarly for cathepsins B, H and L.[15 16] Interestingly the direction of association is such that the alleles that are associated with a higher mRNA level of *CST3* and *CSTA* are associated with lower FEV_1/FVC.

The above findings might seem paradoxical since cystatins are antiproteases and if one simply invoked the protease–antiprotease hypothesis, then one might expect that individuals with higher levels to be protected from COPD. One possibility is that upregulation of *CST3* mRNA and protein could be the result of a feedback loop stimulated by high protease levels. This would be supported by observations in bronchoalveolar lavage fluid and serum reporting higher cystatin C level in patients with emphysema.[17 18] Alternatively excessive inhibition of proteases may confer biological effects on a yet to be discovered mechanism which contributes to the pathogenesis of airflow obstruction. Further molecular studies of the involved proteins might shed more light on this. We previously reported SNPs associated with *CST3* mRNA and protein levels in alveolar macrophages.[19] However, in the latter study, higher *CST3* mRNA in alveolar macrophages was associated with higher FEV_1. Together, these studies suggest that the cystatin genes are important in the pathogenesis of COPD, however the relationship between *CST3* mRNA expression levels and lung function remains to be elucidated in relevant tissues and cell types.

With respect to antioxidant activity, eQTLs from *PPARGC1A* and *GSTO2* were found. *PPARGC1A* binds to peroxisome proliferation-activated receptor γ (PPARγ) by induction of PPARγ ligands, coactivating PPARγ target genes, involved in antioxidant activity. Compared with controls, expression levels of PPARγ, PGC-1α and γ glutamylcysteine synthetase (γ-GCS) have been reported to be significantly increased in the lungs of patients with mild COPD, and progressively decreased in more severe disease.[20] These authors concluded that γ-GCS showed compensatory upregulation in the early stage of COPD, which progressively decompensated with disease progression and that the activation of the PPARγ/PGC-1α pathway may protect against COPD progression by upregulating γ-GCS and relieving oxidative stress. Furthermore *PPARGC1A* has been described to be involved in bronchial smooth muscle remodelling and skeletal muscle wasting.[21 22] For *GSTO2*, a strong association of the Asn142Asp SNP with FEV_1 and FVC was found in the Framingham Heart Study.[23] In a latter study, the Asn142Asp polymorphism in *GSTO2* and the *GSTO1* 140Asp/*GSTO2* 142Asp haplotype were associated with increased risk of COPD but failed to reveal an association between lung function parameters and non-synonymous coding SNPs in the *GSTO* genes.[24] Polymorphisms in *GSTO2* were also associated with COPD either with or without lung cancer.[25]

Pathways involved in apoptosis and in handling of xenobiotic pathways were significantly enriched for causal genes. The results of the present study aid in understanding the underlying genetic dysregulation of apoptosis. In general, increased apoptosis of endothelial cells and fibroblasts has been shown to contribute to the development of emphysema.[26 27] This is partly due to an imbalance caused by excess oxidant and protease effects related to cigarette smoking and to intrinsic dysregulation of apoptosis induced in susceptible individuals. This may explain why the apoptotic effects are larger in smokers with COPD than smokers without COPD. Dysregulation of apoptosis can also work the other way: decreased apoptosis in cells of the immune system can lead to sustained inflammation, and when occurring in fibroblasts (eg, in the bronchial wall) can induce fibrosis. Emphysema is characterised by alveolar cell apoptosis, which was shown to be mediated by increased levels of apoptotic proteins including TRAIL receptors.[28] Interestingly, TRAIL signalling was the most significant pathway enriched with causality genes using Partek. TRAIL signalling has been shown to regulate immune responses in the lung and can lead to a sustained, proinflammatory response that contributes to vascular disease. An opposite effect is related to the death receptor signalling pathway, which, in humans, binds with TRAIL, induces formation of a death-inducing signalling complex, ultimately leading to caspase activation and initiation of apoptosis.[29]

The aryl hydrocarbon receptor signalling pathway is involved in xenobiotic clearance and therefore of relevance to the known vulnerability of patients with COPD to (cigarette) smoke. Furthermore additional metabolic pathways including the glutathione metabolism pathway and two subpathways involving metabolism of the cyanoamino acids, taurine and hypotaurine were also significantly enriched in the causal analysis. Genetic variants in a number of genes in the glutathione pathway have previously been associated with risk for COPD.[25 30]

No other large-scale lung eQTL dataset is available in patients with and without COPD. To replicate the most biologically relevant causality models, we relied on two genome-wide expression datasets. Causality genes were first validated in a gene expression study of bronchial airway epithelial cells obtained by bronchoscopy and then in a transcriptomic study of whole lung explanted at surgery. Whole genome genotyping is not available for these datasets. Accordingly, replication of significant triplets (ie, SNP–mRNA–phenotype) can only evaluate the concordance between gene expression and phenotype. Additional studies with phenotype, genotype and gene expression in the lung would be required to provide full validation.

In conclusion, integration of lung-specific eQTL data with GWAS from the same individuals has revealed interesting and potentially causal pathways of airflow obstruction. GWAS have revolutionised our ability to identify gene variants that contribute to susceptibility for common complex genetic diseases but often do not pinpoint the exact genes or mechanisms. Causal pathway analysis involving the joint examination of genetic and genomic data is a vital next step in discovering novel biomarkers and therapeutic targets in airflow obstruction as evidenced in the present study.

Author affiliations
[1]Institut universitaire de cardiologie et de pneumologie de Québec, Québec, Canada
[2]Department of Pathology and Medical Biology, University of Groningen, University Medical Center Groningen, GRIAC Research Institute, Groningen, The Netherlands
[3]Department of Genetics and Genomics Sciences, Mount Sinai School of Medicine, New York, New York, USA
[4]Department of Molecular Medicine, Laval University, Québec, Canada
[5]Division of Computational Biomedicine, Bioinformatics Program, Boston University, Boston, Massachusetts, USA
[6]University of British Columbia Center for Heart Lung Innovation and Institute for Heart and Lung Health, St Paul's Hospital, Vancouver, British Columbia, Canada
[7]Department of Pathology and Laboratory Medicine, University of British Columbia, Vancouver, British Columbia, Canada
[8]Department of Pulmonology, University of Groningen, University Medical Center Groningen, GRIAC Research Institute, Groningen, The Netherlands
[9]Respiratory Division, Department of Medicine, University of British Columbia, Vancouver, British Columbia, Canada
[10]British Columbia Cancer Agency, Vancouver, British Columbia, Canada
[11]Merck & Co Inc, Rahway, New Jersey, USA

Acknowledgements The authors would like to thank Christine Racine and Sabrina Biardel at the IUCPQ site of the Respiratory Health Network Biobank of the FRQS for their valuable assistance. They also acknowledge the staff at Calcul Québec for IT support with the high-performance computer clusters. At UBC the authors thank the biobank coordinator Dr Mark Elliott. At the Groningen UMCG site Marnix Jonker is thanked for his support in selecting, handling and sending of lung tissues. The authors thank Dr Joshua Millstein of University of Southern California for helpful advice on causality inference methodology.

Contributors WT, KH, YB, MLav, PDP, DDS and DSP were involved in the conception and design of the study. MLam, WT, KH, YB, MLav, KSt, JDC, CC, MC, KSh, JCH, C-AB, SL, MEL, ASp, PDP, DN and DSP were involved in the acquisition and/or the analysis and interpretation of the data. MLam, WT, KH, YB, MB, ASa, PDP, DDS and DSP contributed to the writing or the revision of the manuscript.

Funding This study was funded by Merck Research Laboratories, the Chaire de pneumologie de la Fondation JD Bégin de l'Université Laval, the Fondation de l'Institut universitaire de cardiologie et de pneumologie de Québec, the Respiratory Health Network of the FRQS, the Cancer Research Society and Read for the Cure, and the Canadian Institutes of Health Research (MOP-123369). YB was a research scholar from the Heart and Stroke Foundation of Canada and he is now recipient of a Junior 2 Research Scholar award from the Fonds de recherche Québec—Santé (FRQS). DDS is a Tier 1 Canada Research Chair for COPD. ML is the recipient of a doctoral studentship from the Fonds de recherche Québec - Santé (FRQS).

Competing interests CAB received a grant from Rosetta Merck. DN is a full-time employee of Merck. DSP received consultancy fees from AstraZeneca, Boehringer Ingelheim, Chiesi, GlaxoSmithKline, Takeda, TEVA and a grant from Chiesi. DDS has served on advisory boards of Almirall, Nycomed, Talecris, AstraZeneca, Merck Frosst, Novartis and GlaxoSmithKline, received grants from AstraZeneca, GlaxoSmithKline and Wyeth, and received honoraria for speaking engagements from Takeda, AstraZeneca, GlaxoSmithKline and Boehringer Ingelheim. JDC received consultancy fees from Metera Biosciences and Immuneering. MB received grants from TEVA, AstraZaneca and Chiesi. WT received a grant from Merck Sharp Dohme, received consultancy fees from Pfizer, and received lecture fees from GlaxoSmithKline, Chiesi and Roche Diagnosis.

Patient consent Obtained.

Ethics approval Institutional Review Board guidelines at the three sites.

Provenance and peer review Not commissioned; externally peer reviewed.

REFERENCES

1 Manolio TA. Genomewide association studies and assessment of the risk of disease. *N Engl J Med* 2010;363:166–76.
2 Cho MH, Boutaoui N, Klanderman BJ, *et al*. Variants in FAM13A are associated with chronic obstructive pulmonary disease. *Nat Genet* 2010;42:200–2.
3 Cho MH, Castaldi PJ, Wan ES, *et al*. A genome-wide association study of COPD identifies a susceptibility locus on chromosome 19q13. *Hum Mol Genet* 2012;21:947–57.
4 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.
5 Bosse Y. Updates on the COPD gene list. *Int J Chron Obstruct Pulmon Dis* 2012;7:607–31.
6 Altshuler D, Daly MJ, Lander ES. Genetic mapping in human disease. *Science* 2008;322:881–8.
7 Thun GA, Imboden M, Ferrarotti I, *et al*. Causal and synthetic associations of variants in the SERPINA gene cluster with alpha1-antitrypsin serum levels. *PLoS Genet* 2013;9:e1003585.
8 Nicolae DL, Gamazon E, Zhang W, *et al*. Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. *PLoS Genet* 2010;6:e1000888.
9 Hao K, Bosse Y, Nickle DC, *et al*. Lung eQTLs to help reveal the molecular underpinnings of asthma. *PLoS Genet* 2012;8:e1003029.
10 Lamontagne M, Couture C, Postma DS, *et al*. Refining susceptibility loci of chronic obstructive pulmonary disease with lung eQTLs. *PLoS ONE* 2013;8:e70220.
11 Sieberts SK, Schadt EE. Moving toward a system genetics view of disease. *Mamm Genome* 2007;18:389–401.
12 Schadt EE, Lamb J, Yang X, *et al*. An integrative genomics approach to infer causal associations between gene expression and disease. *Nat Genet* 2005;37:710–7.
13 Steiling K, van den Berge M, Hijazi K, *et al*. A dynamic bronchial airway gene expression signature of chronic obstructive pulmonary disease and lung function impairment. *Am J Respir Crit Care Med* 2013;187:933–42.
14 Campbell JD, McDonough JE, Zeskind JE, *et al*. A gene expression signature of emphysema-related lung destruction and its reversal by the tripeptide GHK. *Genome Med* 2012;4:67.
15 Butler MW, Fukui T, Salit J, *et al*. Modulation of cystatin A expression in human airway epithelium related to genotype, smoking, COPD, and lung cancer. *Cancer Res* 2011;71:2572–81.
16 Pavlova A, Bjork I. Grafting of features of cystatins C or B into the N-terminal region or second binding loop of cystatin A (stefin A) substantially enhances inhibition of cysteine proteinases. *Biochemistry* 2003;42:11326–33.
17 Rokadia HK, Agarwal S. Serum cystatin C and emphysema: results from the National Health and Nutrition Examination Survey (NHANES). *Lung* 2012;190:283–90.
18 Takeyabu K, Betsuyaku T, Nishimura M, *et al*. Cysteine proteinases and cystatin C in bronchoalveolar lavage fluid from subjects with subclinical emphysema. *Eur Respir J* 1998;12:1033–9.
19 Ishii T, Abboud RT, Wallace AM, *et al*. Alveolar macrophage proteinase/antiproteinase expression and lung function/emphysema. *Eur Respir J* 2014;43:82–91.
20 Li J, Dai A, Hu R, *et al*. Positive correlation between PPARgamma/PGC-1alpha and gamma-GCS in lungs of rats and patients with chronic obstructive pulmonary disease. *Acta Biochim Biophys Sin (Shanghai)* 2010;42:603–14.
21 Trian T, Benard G, Begueret H, *et al*. Bronchial smooth muscle remodeling involves calcium-dependent enhanced mitochondrial biogenesis in asthma. *J Exp Med* 2007;204:3173–81.
22 Remels AH, Gosker HR, Schrauwen P, *et al*. TNF-alpha impairs regulation of muscle oxidative phenotype: implications for cachexia? *FASEB J* 2010;24:5052–62.

23 Wilk JB, Walter RE, Laramie JM, *et al*. Framingham Heart Study genome-wide association: results for pulmonary function measures. *BMC Med Genet* 2007;8(Suppl 1):S8.
24 Yanbaeva DG, Wouters EF, Dentener MA, *et al*. Association of glutathione-S-transferase omega haplotypes with susceptibility to chronic obstructive pulmonary disease. *Free Radic Res* 2009;43:738–43.
25 de Andrade M, Li Y, Marks RS, *et al*. Genetic variants associated with the risk of chronic obstructive pulmonary disease with and without lung cancer. *Cancer Prev Res (Phila)* 2012;5:365–73.
26 Park JW, Ryter SW, Kyung SY, *et al*. The phosphodiesterase 4 inhibitor rolipram protects against cigarette smoke extract-induced apoptosis in human lung fibroblasts. *Eur J Pharmacol* 2013;706:76–83.
27 Yang Q, Underwood MJ, Hsin MK, *et al*. Dysfunction of pulmonary vascular endothelium in chronic obstructive pulmonary disease: basic considerations for future drug development. *Curr Drug Metab* 2008;9:661–7.
28 Morissette MC, Vachon-Beaudoin G, Parent J, *et al*. Increased p53 level, Bax/Bcl-x (L) ratio, and TRAIL receptor expression in human emphysema. *Am J Respir Crit Care Med* 2008;178:240–7.
29 Bodmer JL, Holler N, Reynard S, *et al*. TRAIL receptor-2 signals apoptosis through FADD and caspase-8. *Nat Cell Biol* 2000;2:241–3.
30 He JQ, Connett JE, Anthonisen NR, *et al*. Glutathione S-transferase variants and their interaction with smoking on lung function. *Am J Respir Crit Care Med* 2004;170:388–94.

Online Data Supplement

Genetic regulation of gene expression in the lung identifies CST3 and CD22 as potential causal genes for airflow obstruction

Maxime Lamontagne*[1], Wim Timens*[2], Ke Hao*[3], Yohan Bossé[1,4], Michel Laviolette[1], Katrina Steiling[5], Joshua D Campbell[5], Christian Couture[1], Massimo Conti[1], Karen Sherwood[6], James C. Hogg[6,7], Corry-Anke Brandsma[2], Maarten van den Berge[8], Andrew Sandford[6,9], Stephen Lam[10], Marc E Lenburg[5], Avrum Spira[5], Peter D Paré[6,9], David Nickle[11], Don D. Sin[6,9], Dirkje S. Postma[†8]

Methods

Subject selection

The methods for subject selection and phenotyping as well as for interrogation of gene expression and genotype were recently described[1]. The lung tissue used for discovery of eQTLs was from 1,111 human subjects who underwent lung surgery at three academic sites, Laval University, University of British Columbia (UBC), and University of Groningen, henceforth referred to as Laval, UBC, and Groningen, respectively. For the present study, the principal aim was to examine smoking-related airway obstruction. Thus, we excluded subjects whose lung function may have been influenced by lung diseases other than COPD. Exclusion criteria were: 1) missing data for lung function, i.e. for both Forced Expiratory Volume in 1 second as a percentage of its predicted value (FEV_1 % predicted) and that divided by Forced Vital Capacity (FEV_1/FVC) and 2) patients with asthma, extensive pneumonia, cystic fibrosis, bronchiectasis, pulmonary fibrosis, pulmonary hypertension, primary bullous emphysema, mesothelioma, diffuse alveolar damage and alpha-1-antitrypsin deficiency. We did not exclude individuals with lung cancer, focal pneumonia or atelectasis, but in these cases lung tissue samples were taken as far away as possible from involved areas. Although lung function in theory could be compromised in these cases, the vast majority of subjects included in the biobanks from the three participating sites had

relatively small, non-obstructing lung tumors which were unlikely to significantly interfere with the subjects' lung function. Of the 1,111 subjects in whom there were data on genotype and gene expression, 848 with sufficient phenotypic information and without a lung disease (other than COPD and lung cancer) were included in the analysis. The demographic and clinical features of the subjects in the three cohorts are described in **Table 1**.

COPD phenotypes

For the primary analyses, three phenotypes were used: FEV_1 % predicted and FEV_1/FVC as continuous variables, and COPD defined dichotomously based on an FEV_1/FVC < 0.7 cutoff. We used post-bronchodilator spirometry when available; otherwise, pre-bronchodilator values were used.

Genome-wide association study

Genome-wide association was performed on the three phenotypes using linear or logistic regression models. Single-marker association tests were run within each cohort adjusting for age, gender and smoking status. Furthermore, we conducted a fixed-effects meta-analysis combining the three cohorts using inverse standard

error weighting. We performed genomic control (GC) correction for individual cohorts and for the meta-analysis. To avoid over-correction, we computed the genomic inflation factor (□) on the 90% of SNPs with the largest p-values. The □ estimates were small (≤1.03) for the individual cohort analyses and the meta-analysis.

Causality Models

We evaluated three competing causality models to describe the relationship between lung eQTL-SNPs, RNA expression and COPD phenotypes (**Figure 1**). We tested three linear equations in describing the molecular relationships,

$$T_i = \alpha_1 + \beta_1 L_i + \varepsilon_{1i} \quad \text{(eq 1)}$$

$$G_i = \alpha_2 + \beta_2 T_i + \beta_3 L_i + \varepsilon_{2i} \quad \text{(eq 2)}$$

$$T_i = \alpha_3 + \beta_4 G_i + \beta_5 L_i + \varepsilon_{3i} \quad \text{(eq 3)}$$

A SNP at a specific locus is denoted by L, gene expression for a specific transcript by G, and a measured clinical endpoint by T (i.e. FEV_1 % predicted, FEV_1/FVC or COPD yes/no). Dependencies are likely to exist between certain pairs of covariates in the preceding three models. We inferred a causal relationship using conditional correlation[2]. We required a p-value for $\beta_1 < 1 \times 10^{-3}$ before examining a particular (L, G, T) variable triplet. If a p-value for $\beta_3 < 0.05$ and p-value for $\beta_5 > 0.05$, the causal model was selected. If the converse was true, a reactive model was

selected. If both p-values were < 0.05, an independence model was selected. If both p-values were > 0.05 then no call was made. Model 1 indicates a potential causal relationship and was the main interest of this work.

Analyses were performed in the three cohorts separately and then combined into a meta-analysis. We then conducted sample bootstrapping and repeated the causality test for N=1000 realizations. The reliability score of the molecular relationship is the fraction of bootstrap realizations that supports the call observed.

Manual curation of significant and reliable causal models

To explore the biology and possible role of genes in the causal model, experts in the pathobiology of COPD at each of the participating sites manually reviewed the literature for functionality and known genetic associations. Several strategies were used including PubMed searches using the gene's name and COPD or emphysema, interrogation of NCBI PubMed Gene site, GeneRIFs (Gene Reference Into Functions database), PheGenI for eQTL, and phenotype association data.

Pathway analyses on causality results

Pathway analyses were performed on causality results in order to search for enrichments of specific gene pathways. Genes that fitted Model 1 (causal) with

reliability scores > 0.8 were analyzed using the Ingenuity Pathways Analysis (IPA, Ingenuity® Systems, www.ingenuity com). Causal genes were mapped to corresponding gene objects in IPA using official gene symbols and overlaid onto canonical pathways contained in the Ingenuity Pathways Knowledge Base. The latter analysis identified pathways from the IPA library of canonical pathways that were most enriched with causal genes. The significance of the association with canonical pathways was determined using a right-tailed Fisher's exact test. This test compared the number of causal genes versus total genes in a canonical pathway beyond that expected by chance alone. A total of 181 canonical pathways were tested. A nominal p-value < 0.05 was considered significant. Similarly the same list of causality genes was examined for enrichment in gene ontology and functional categories using MetaCore™ (version 6.12, build 42289, GeneGo, Inc.) and mapped to both the Gene Ontology (GO) project and proprietary ontologies in the MetaCore™ knowledge database. The significance of functional enrichment of genes was determined by using a False Discovery Rate (FDR) cutoff of 0.05. Network analysis was also carried out using canonical networks in the Metacore™ knowledge database. Direct interaction networks based on our gene lists were built using seed nodes and their direct interactions were assessed using curated, known interactions. Pathway analysis was also carried out using Partek Genomic Suite 6.6 software (Partek GS, version 6.12.0907, www.partek.com). Causality genes were

mapped to official gene symbols in Partek GS and parsed to Partek Pathway software. Partek Pathway interrogates the REACTOME (www.reactome.org) and KEGG (www.genome.jp/kegg/) databases and identifies canonical pathways that are enriched with causality genes. Enrichment score p-values were derived using Fisher's Exact test. A p-value < 0.05 was considered significant.

Replication studies

Bronchial airway epithelial dataset. Bronchial airway brushings were obtained during bronchoscopy from active and former smokers who were enrolled in a lung cancer screening program at the British Columbia Cancer Research Agency[3]. Institutional Review Board approval was obtained at participating institutions, and all subjects provided written informed consent. RNA isolated from bronchial brushings of 238 lung cancer-free active and former smokers with and without COPD was profiled using Affymetrix Human Gene 1.0 ST Arrays (Affymetrix Inc., Santa Clara, CA, USA). Microarray data were deposited in the Gene Expression Omnibus (GSE37147). Gene expression estimates were derived and normalized as previously described[3]. Gene expression levels associated with FEV_1 % predicted, FEV_1/FVC and the presence of COPD were determined using linear modeling as previously described[3].

Regional lung tissue dataset. Whole lungs were explanted from patients with severe COPD (n=6) and from donors (n=2). Each lung was sampled in eight consecutive regions from the apex to the base of the lung. This study was approved by the institutional review board and written informed consent was obtained from each patient prior to surgery or from the next of kin of the persons who the donated lung. The gene expression profiles were obtained in 64 samples (8 patients x 8 regions) using the Human Exon 1.0 ST array. The gene expression dataset was deposited in the Gene Expression Omnibus (GSE27597). Details of sample collection and processing as well as analytical methods to normalize and obtain gene expression values have been previously described[4]. The degree of emphysema in each sample was quantified by measuring the mean linear intercept (Lm) on micro-computed tomography (CT) scans for tissues. Lm represents a morphological measurement of alveolar destruction and is a surrogate for emphysema severity. In this study, genes were associated with regional emphysema severity within the same lung using linear models as described before[4].

Table S2. Models of biological relevance identified by manual curation

SNPs	Gene (probe set)	Gene role & references
FEV1 % predicted		
rs769178	*NCR3* (100125842_TGI_at)	SNPs near *NCR3* gene were associated with lung function[5].
rs6048956	*CST3* (100307577_TGI_at)	Protease-antiprotease balance[6].
rs6515375	*CST3* (100125967_TGI_at)	Protease-antiprotease balance[6].
rs2270859	*CSTA* (100148334_TGI_at)	Expression of CSTA is modulated by genotype, smoking, COPD and lung cancer[7].
rs4550905	*PPARGC1A* (100131093_TGI_at)	The PPAR□-PGC-1α pathway regulates antioxidant genes and protects against COPD in rats[8].
rs3803761	*FLCN* (100135396_TGI_at)	Mutations in *FLCN* cause Birt-Hogg-Dubé syndrome, a monogenic disorder characterized by spontaneous pneumothorax. Genetic variants in *FLCN* were not associated with severe, early-onset COPD[9].
rs1543438	*BCL2L1* (100158784_TGI_at)	Cigarette smoke extract induces the expression of BCL2L1in human dendritic cells and augments survival of these cells in COPD patients[10]
rs9880397	*CADM2* (100162763_TGI_at)	SNPs in *CADM2* were associated with lung function and asthma (unpublished).
rs2466183	TNFRSF10B (100153254_TGI_at)	Involved in T cell and eosinophil regulation in bronchial smooth muscle cell death in asthma[11]. Increased expression of the protein in the lung of subjects with

		emphysema[12].
FEV_1/FVC		
rs17754977	*GSTO2* (100132911_TGI_at)	GSTO enzymes protect against oxidative stress. Involved in the biotransformation of arsenic found in cigarette smoke[13]. SNPs in *GSTO2* were associated with FEV_1, FVC[14] and COPD[15].
rs9987135	*DEPDC6* (100154484_TGI_at)	DEPDC6 is a negative regulator of mTORC1 and mTORC2 signaling pathways. Decreased TOR activity has been found to slow aging in yeast, worms, flies, and mice[16]. The mTOR inhibitor, rapamycin, increases the lifespan in mice and reduces bronchial hyperresponsiveness and airway remodelling[17].
rs10411704	*CD22* (100154732_TGI_at)	CD22 is present on the surface of (mature) B cells and prevents over activation of the immune system and development of autoimmune system.
rs12179536	*MUC22* (100304000_TGI_at)	Polymorphisms in *MUC22* were associated with panbronchiolitis[18]. MUC22 is expressed in the lung.
rs2287765	*SPINK5* (100305138_TGI_at)	*SPINK5* is a candidate gene for asthma and allergy[19]. Upregulation of SPINK5 in epithelial cell line increases inflammatory responses[20].

REFERENCES

1. Hao K, Bosse Y, Nickle DC, et al. Lung eQTLs to Help Reveal the Molecular Underpinnings of Asthma. *PLoS Genet* 2012;8:e1003029.

2. Schadt EE, Lamb J, Yang X, et al. An integrative genomics approach to infer causal associations between gene expression and disease. *Nat Genet* 2005;37:710-7.

3. Steiling K, van den Berge M, Hijazi K, et al. A dynamic bronchial airway gene expression signature of chronic obstructive pulmonary disease and lung function impairment. *Am J Respir Crit Care Med* 2013;187:933-42.

4. Campbell JD, McDonough JE, Zeskind JE, et al. A gene expression signature of emphysema-related lung destruction and its reversal by the tripeptide GHK. *Genome medicine* 2012;4:67.

5. Soler Artigas M, Loth DW, Wain LV, et al. Genome-wide association and large-scale follow up identifies 16 new loci influencing lung function. *Nat Genet* 2011;43:1082-90.

6. Abboud RT, Vimalanathan S. Pathogenesis of COPD. Part I. The role of protease-antiprotease imbalance in emphysema. *Int J Tuberc Lung Dis* 2008;12:361-7.

7. Butler MW, Fukui T, Salit J, et al. Modulation of cystatin A expression in human airway epithelium related to genotype, smoking, COPD, and lung cancer. *Cancer Res* 2011;71:2572-81.

8. Li J, Dai A, Hu R, et al. Positive correlation between PPARgamma/PGC-1alpha and gamma-GCS in lungs of rats and patients with chronic obstructive pulmonary disease. *Acta Biochim Biophys Sin (Shanghai)* 2010;42:603-14.

9. Cho MH, Klanderman BJ, Litonjua AA, et al. Folliculin mutations are not associated with severe COPD. *BMC Med Genet* 2008;9:120.

10. Vassallo R, Walters PR, Lamont J, et al. Cigarette smoke promotes dendritic cell accumulation in COPD; a Lung Tissue Research Consortium study. *Respir Res* 2010;11:45.

11. Solarewicz-Madejek K, Basinski TM, Crameri R, et al. T cells and eosinophils in bronchial smooth muscle cell death in asthma. *Clin Exp Allergy* 2009;39:845-55.

12. Morissette MC, Vachon-Beaudoin G, Parent J, et al. Increased p53 level, Bax/Bcl-x(L) ratio, and TRAIL receptor expression in human emphysema. *Am J Respir Crit Care Med* 2008;178:240-7.

13. Lesseur C, Gilbert-Diamond D, Andrew AS, et al. A case-control study of polymorphisms in xenobiotic and arsenic metabolism genes and arsenic-related bladder cancer in New Hampshire. *Toxicol Lett* 2012;210:100-6.

14. Wilk JB, Walter RE, Laramie JM, et al. Framingham Heart Study genome-wide association: results for pulmonary function measures. *BMC Med Genet* 2007;8 Suppl 1:S8.

15. Yanbaeva DG, Wouters EF, Dentener MA, et al. Association of glutathione-S-transferase omega haplotypes with susceptibility to chronic obstructive pulmonary disease. *Free Radic Res* 2009;43:738-43.

16. Evans DS, Kapahi P, Hsueh WC, et al. TOR signaling never gets old: aging, longevity and TORC1 activity. *Ageing Res Rev* 2011;10:225-37.

17. Kramer EL, Hardie WD, Mushaben EM, et al. Rapamycin decreases airway remodeling and hyperreactivity in a transgenic model of noninflammatory lung disease. *J Appl Physiol* 2011;111:1760-7.

18. Hijikata M, Matsushita I, Tanaka G, et al. Molecular cloning of two novel mucin-like genes in the disease-susceptibility locus for diffuse panbronchiolitis. *Hum Genet* 2011;129:117-28.

19. Kabesch M, Carr D, Weiland SK, et al. Association between polymorphisms in serine protease inhibitor, kazal type 5 and asthma phenotypes in a large German population sample. *Clin Exp Allergy* 2004;34:340-5.

20. Birben E, Sackesen C, Turgutoglu N, et al. The role of SPINK5 in asthma related physiological events in the airway epithelium. *Respir Med* 2012;106:349-55.

Supplementary Figures

Figure S1. Direction of effects for causality model with rs6515375, mRNA expression of CST3, and FEV1 % predicted. Data are presented as described in Figure 2.

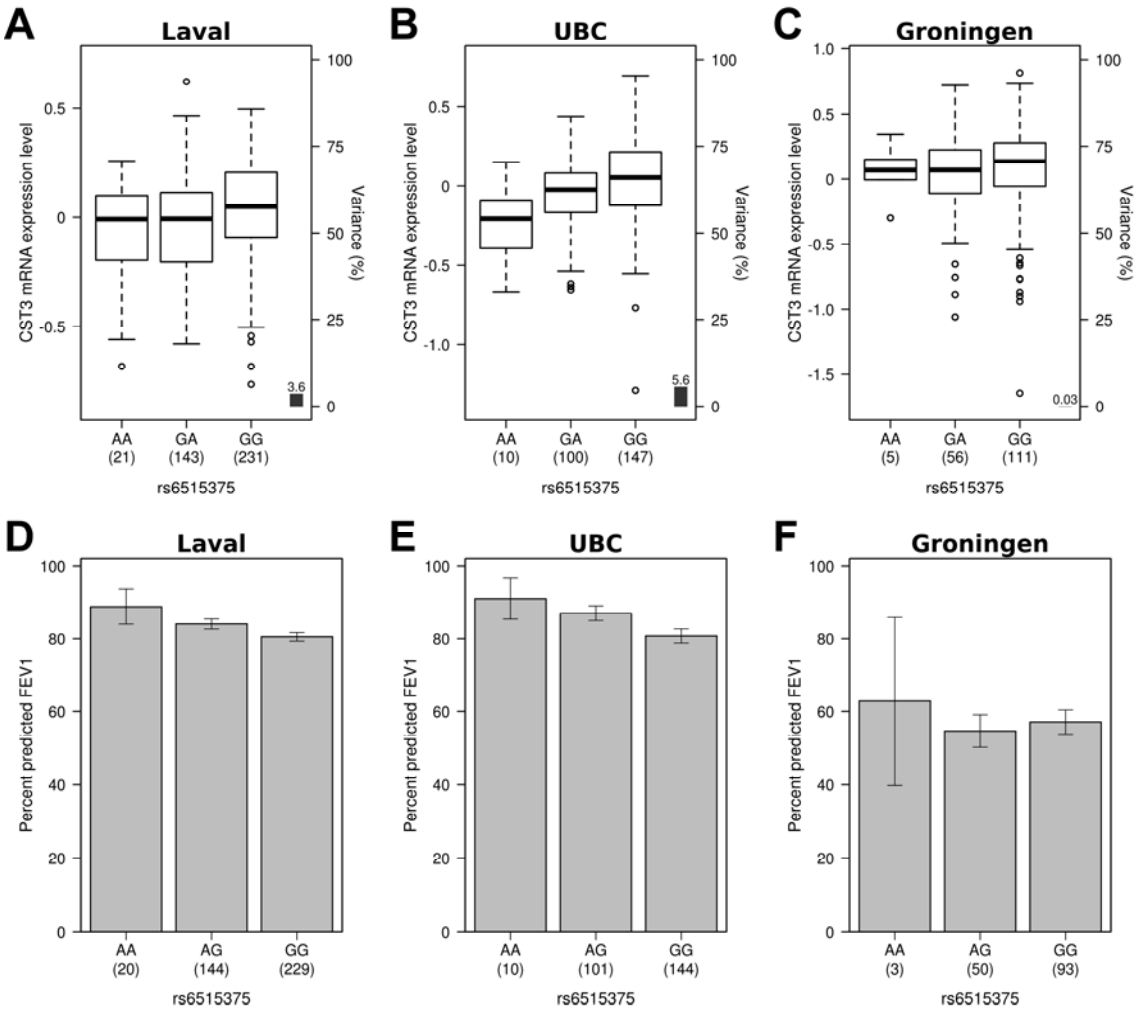


Figure S2. Direction of effects for causality model with rs769178, mRNA expression of NCR3, and FEV1 % predicted. Data are presented as described in Figure 2.

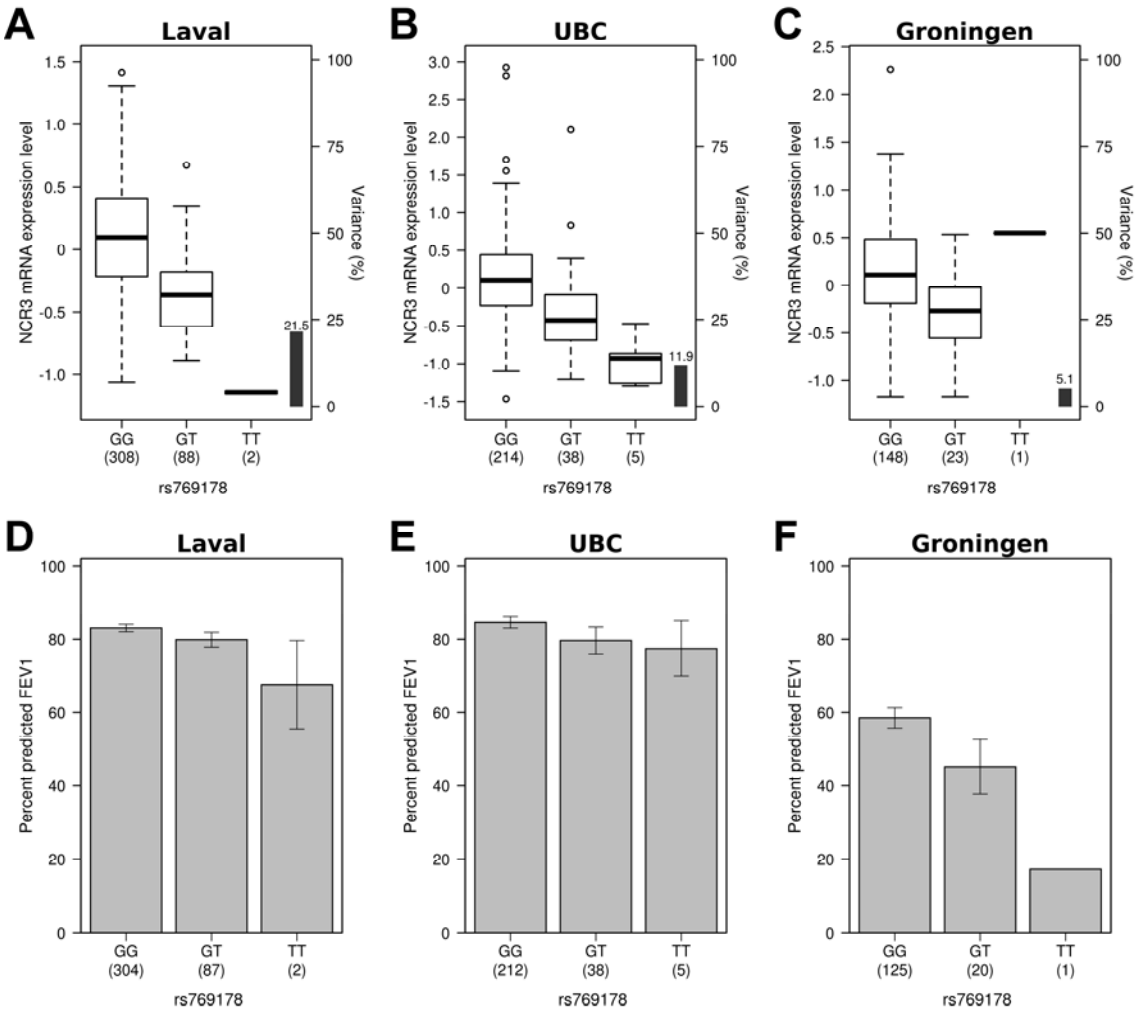


Figure S3. Direction of effects for causality model with rs2270859, mRNA expression of CSTA, and FEV1 % predicted. Data are presented as described in Figure 2.

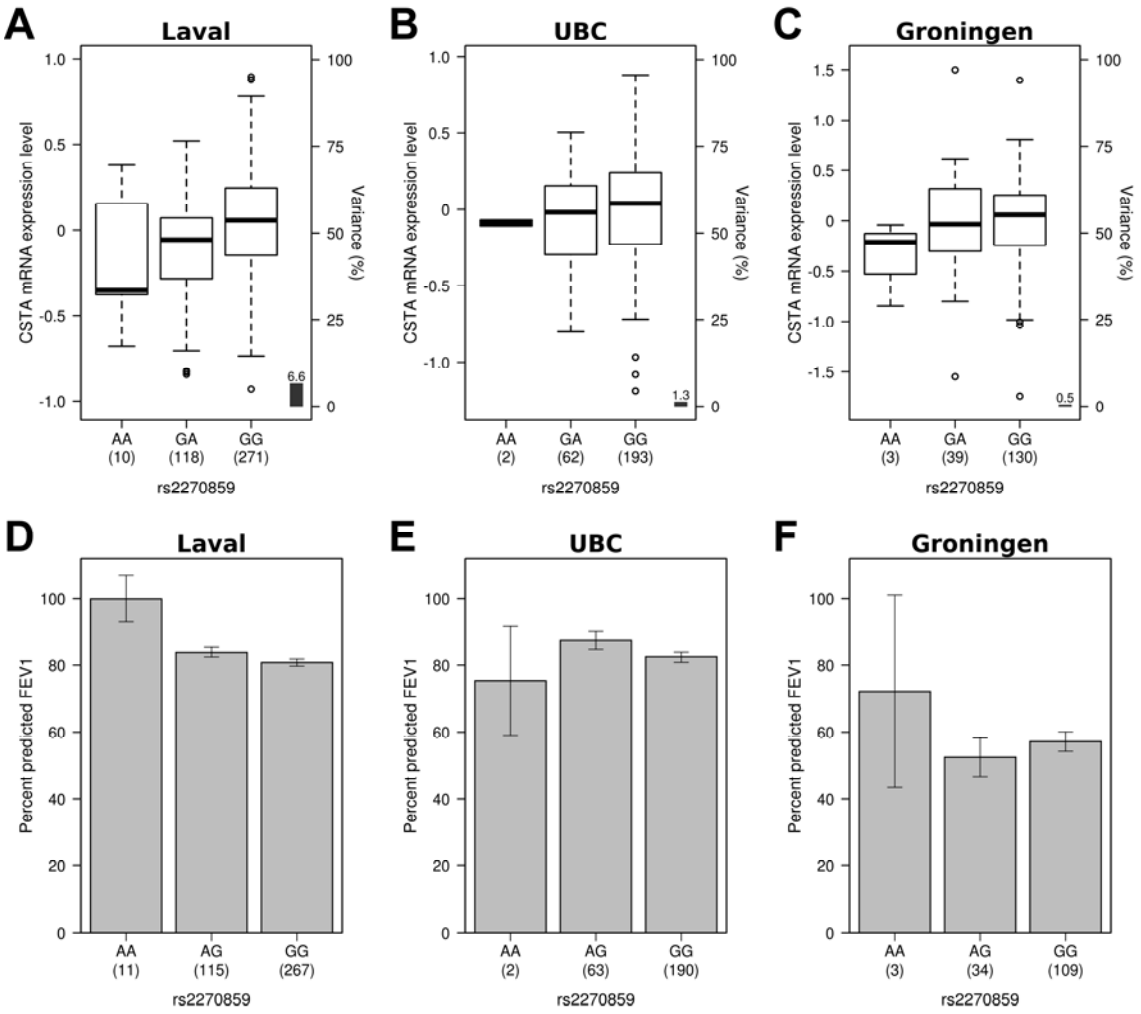


Figure S4. Direction of effects for causality model with rs4550905, mRNA expression of PPARGC1A, and FEV1 % predicted. Data are presented as described in Figure 2.

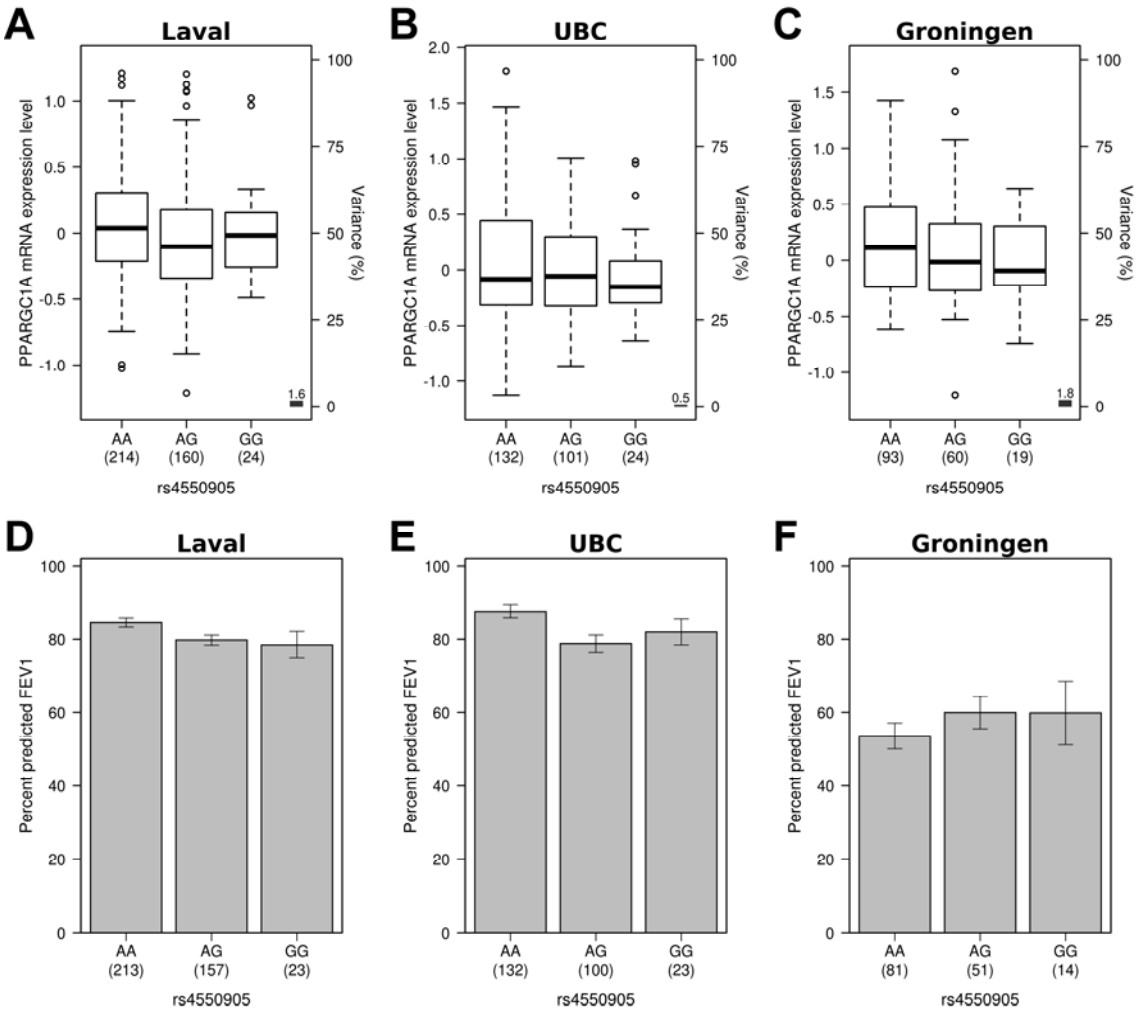


Figure S5. Direction of effects for causality model with rs3803761, mRNA expression of FLCN, and FEV1 % predicted. Data are presented as described in Figure 2.

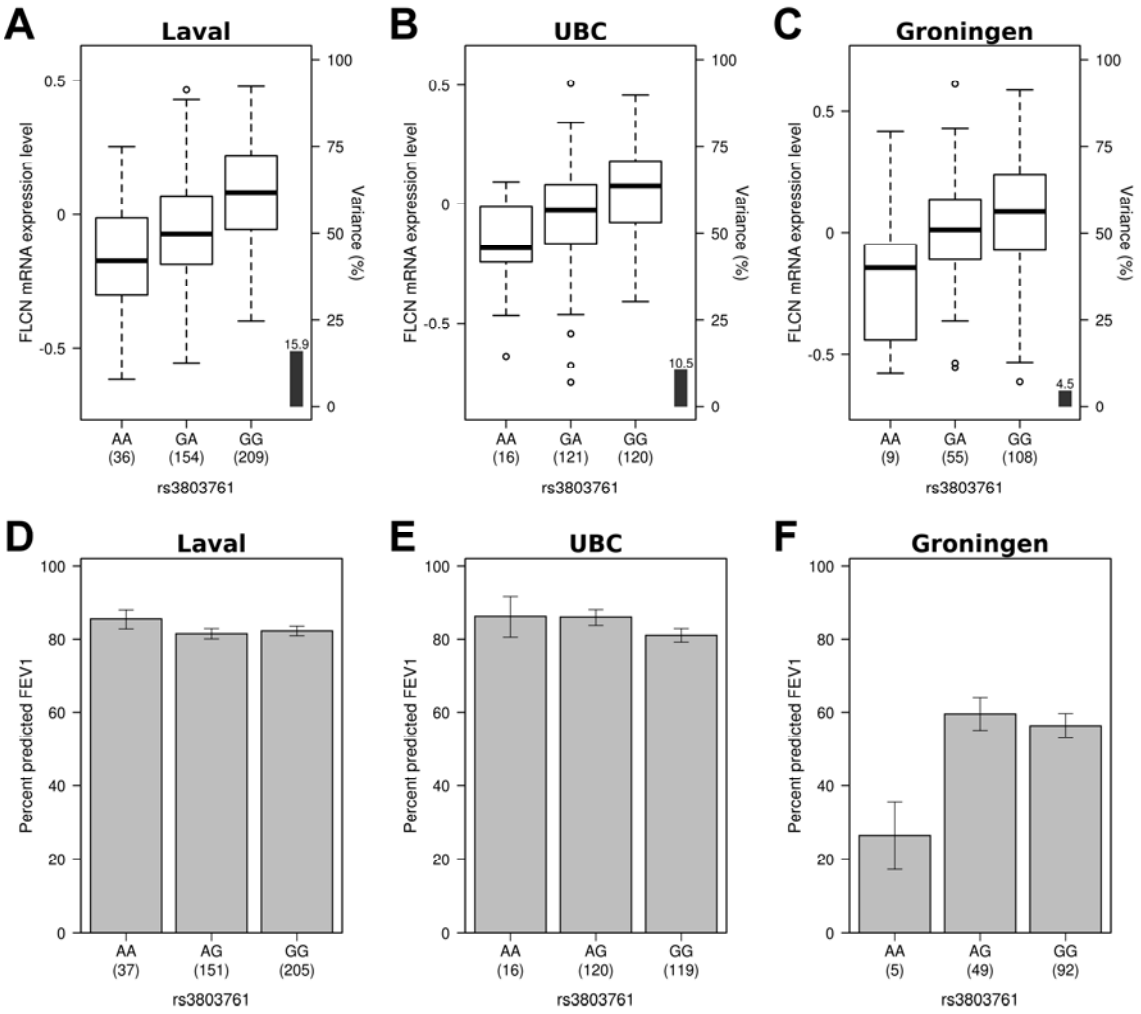


Figure S6. Direction of effects for causality model with rs1543438, mRNA expression of BCL2L1, and FEV1 % predicted. Data are presented as described in Figure 2.

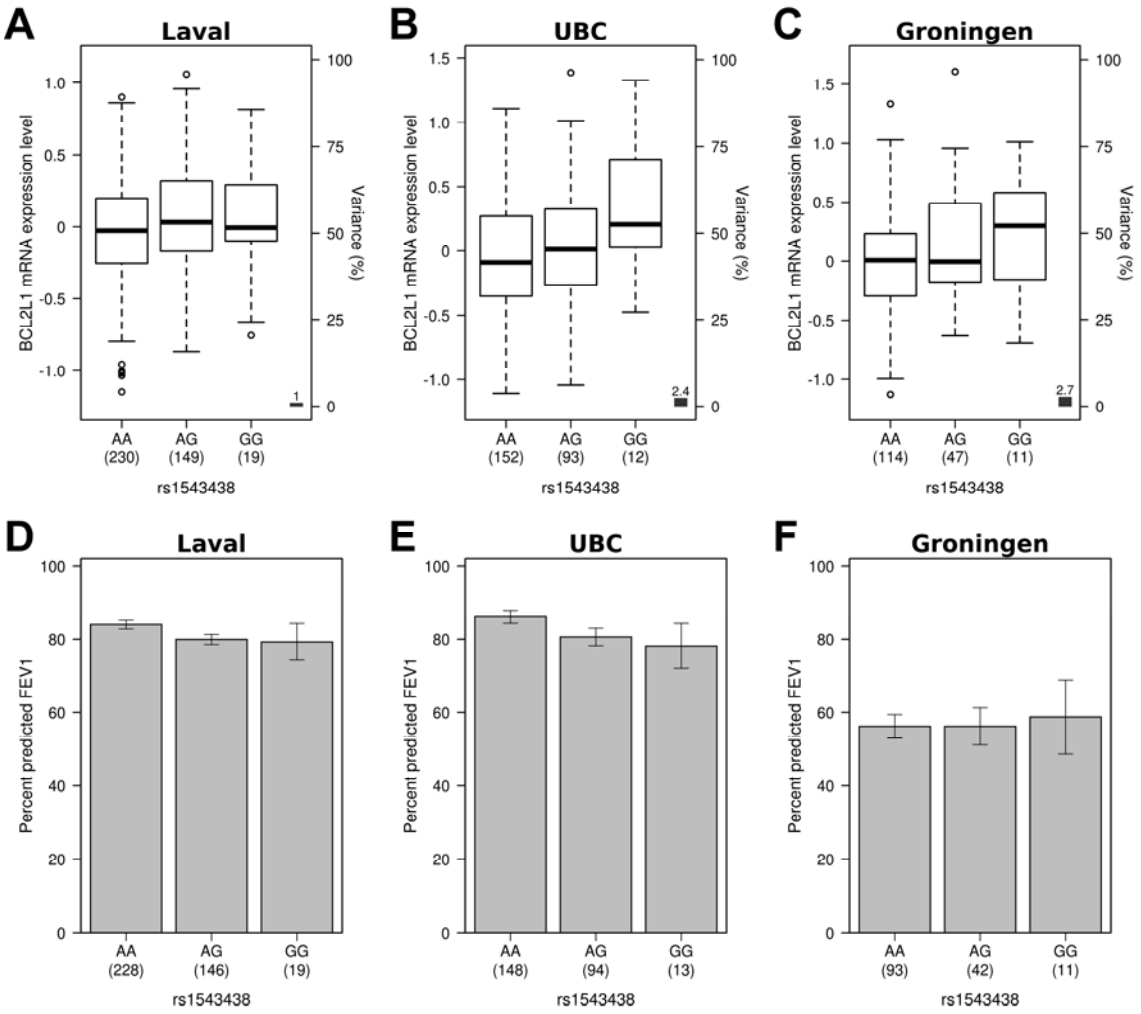


Figure S7. Direction of effects for causality model with rs9880397, mRNA expression of CADM2, and FEV1 % predicted. Data are presented as described in Figure 2.

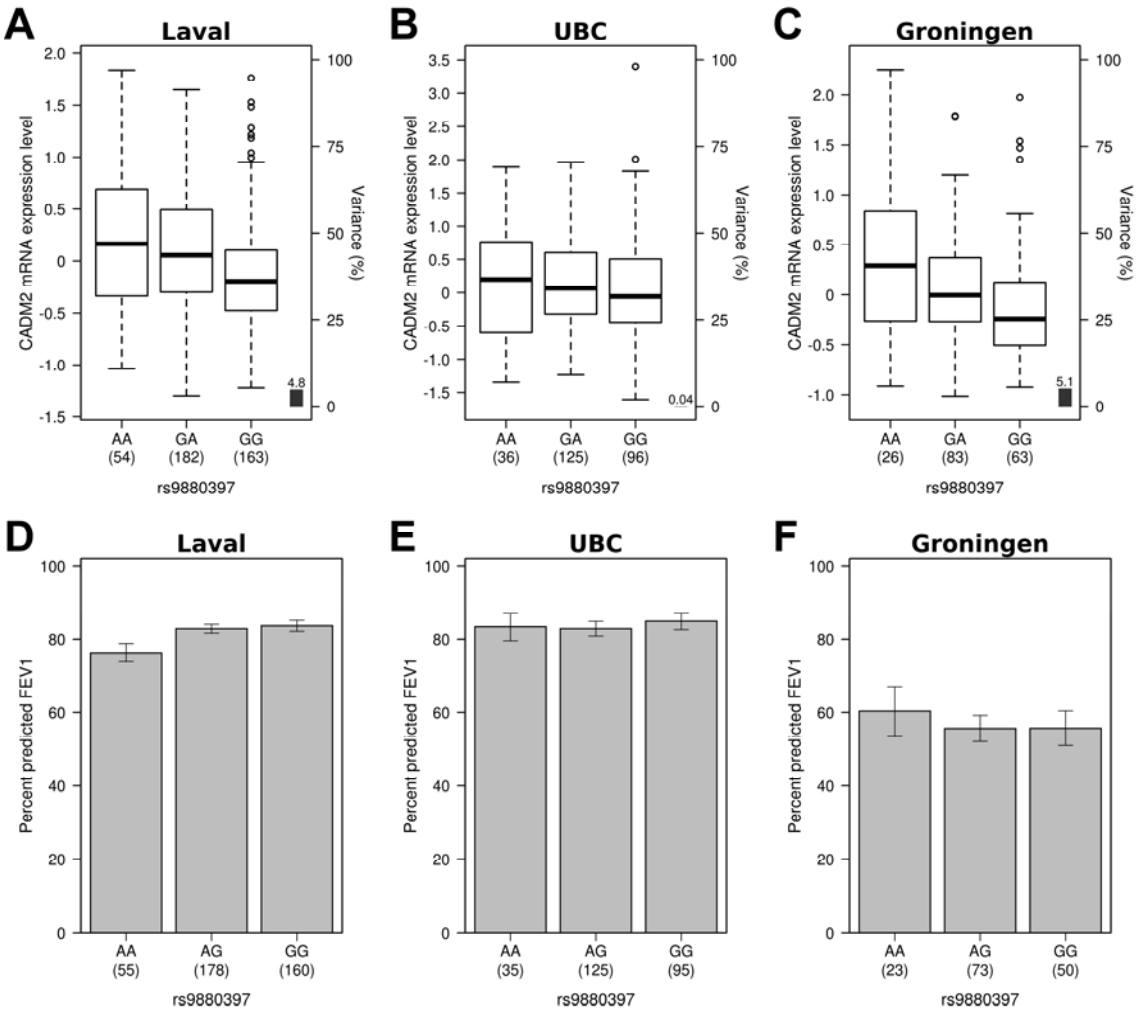


Figure S8. Direction of effects for causality model with rs2466183, mRNA expression of TNFRSF10B, and FEV1 % predicted. Data are presented as described in Figure 2.

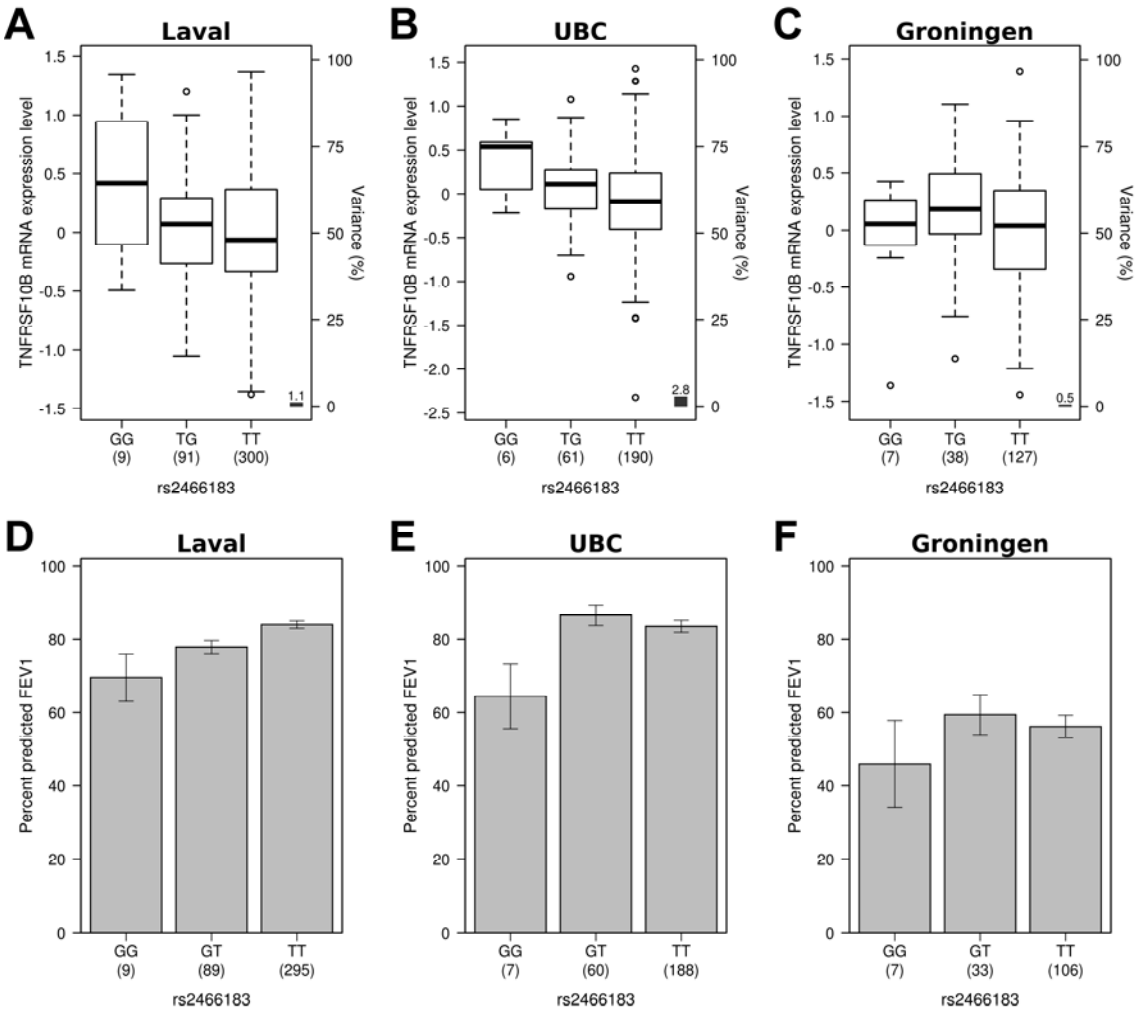


Figure S9. Direction of effects for causality model with rs17754977, mRNA expression of GSTO2, and FEV1/FVC. Data are presented as described in Figure 2.

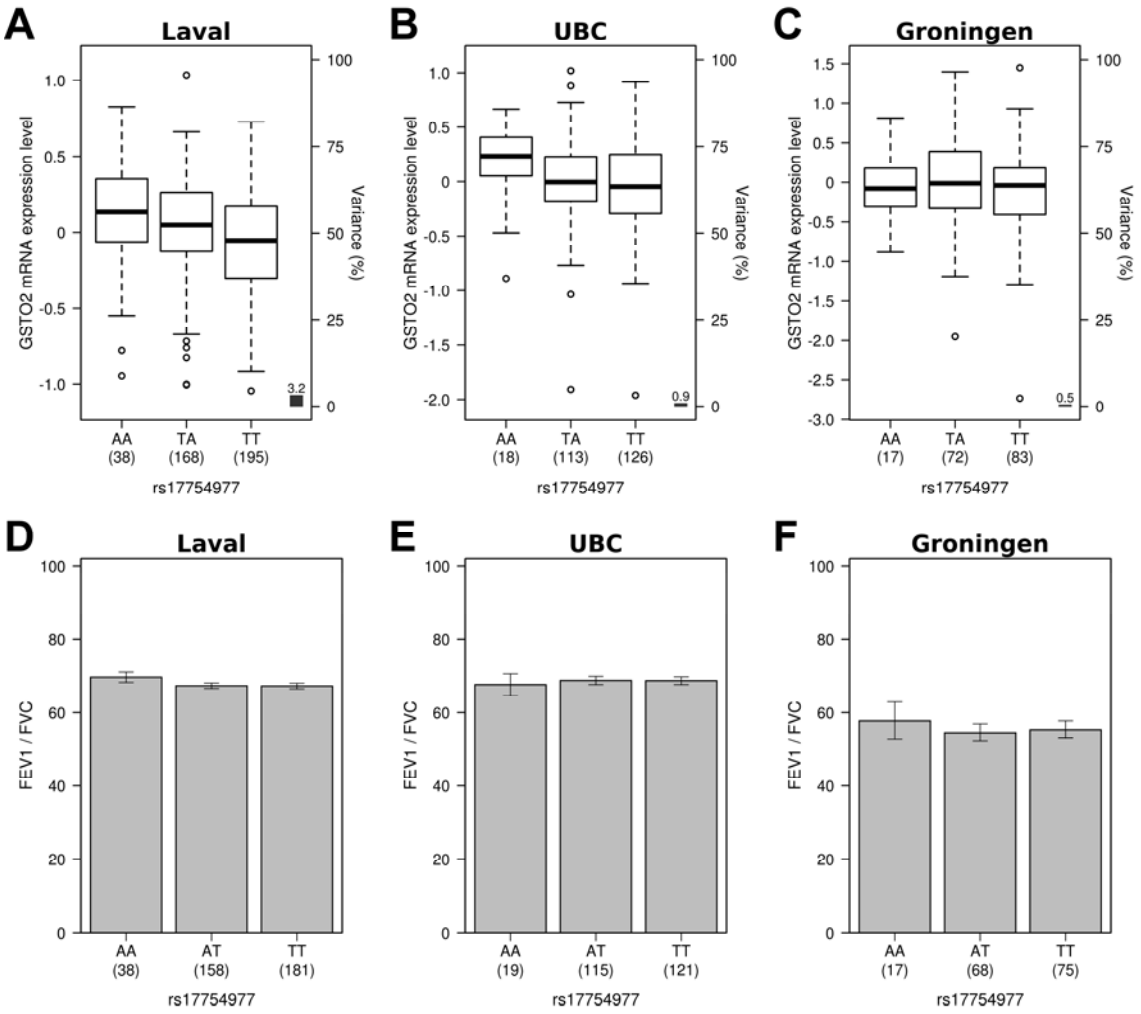


Figure S10. Direction of effects for causality model with rs9987135, mRNA expression of DEPDC6, and FEV1/FVC. Data are presented as described in Figure 2.

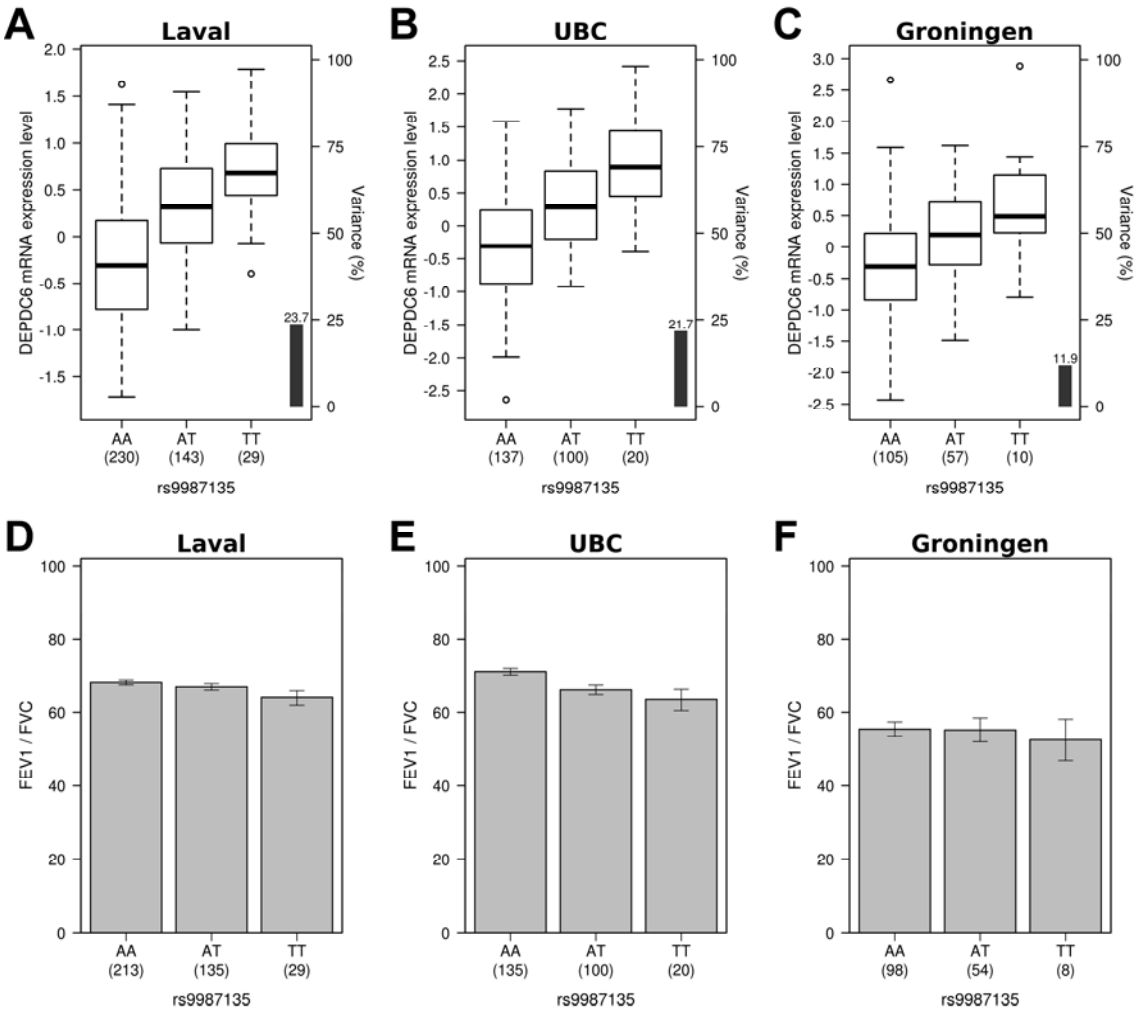


Figure S11. Direction of effects for causality model with rs10411704, mRNA expression of CD22, and FEV1/FVC. Data are presented as described in Figure 2.

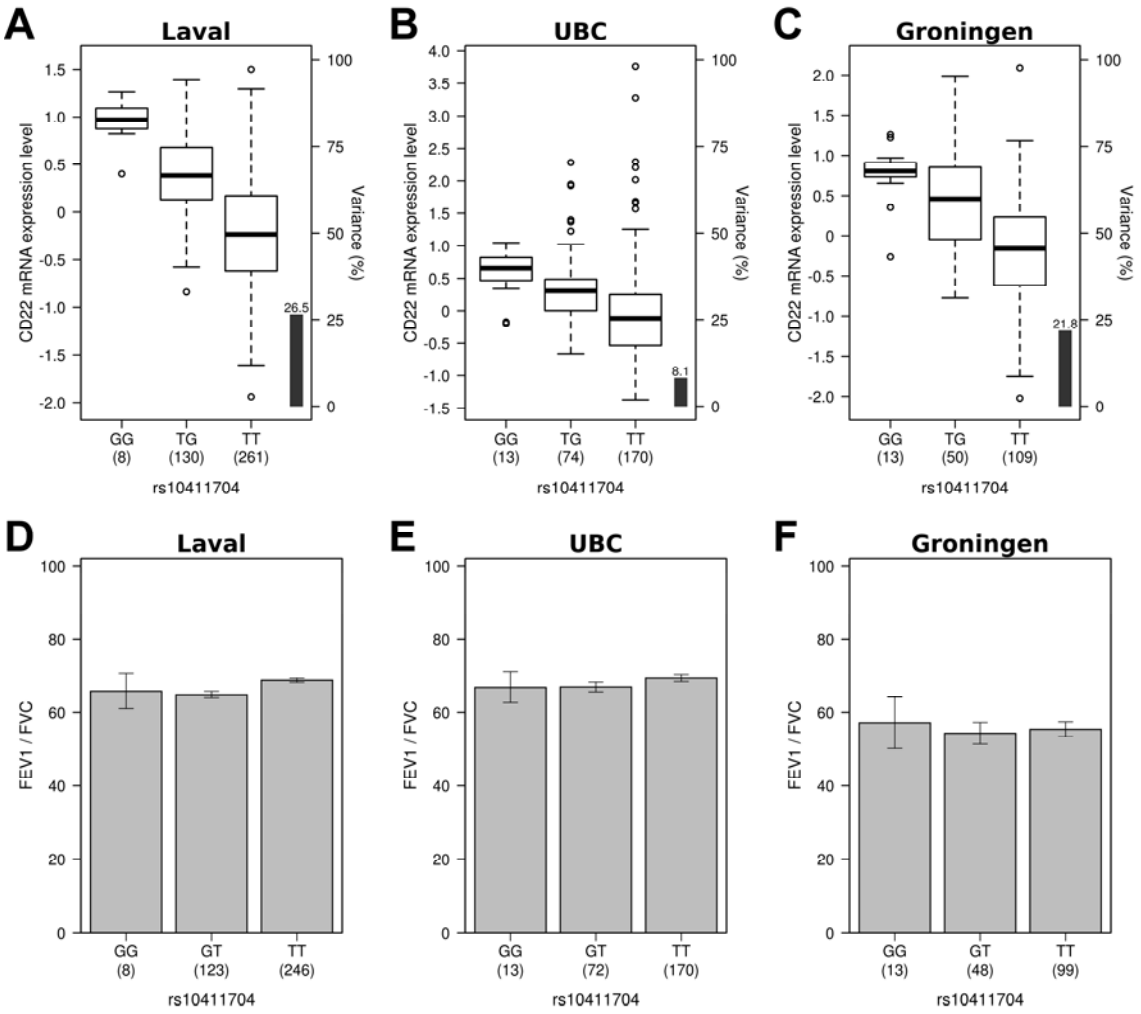


Figure S12. Direction of effects for causality model with rs12179536, mRNA expression of MUC22, and FEV1/FVC. Data are presented as described in Figure 2.

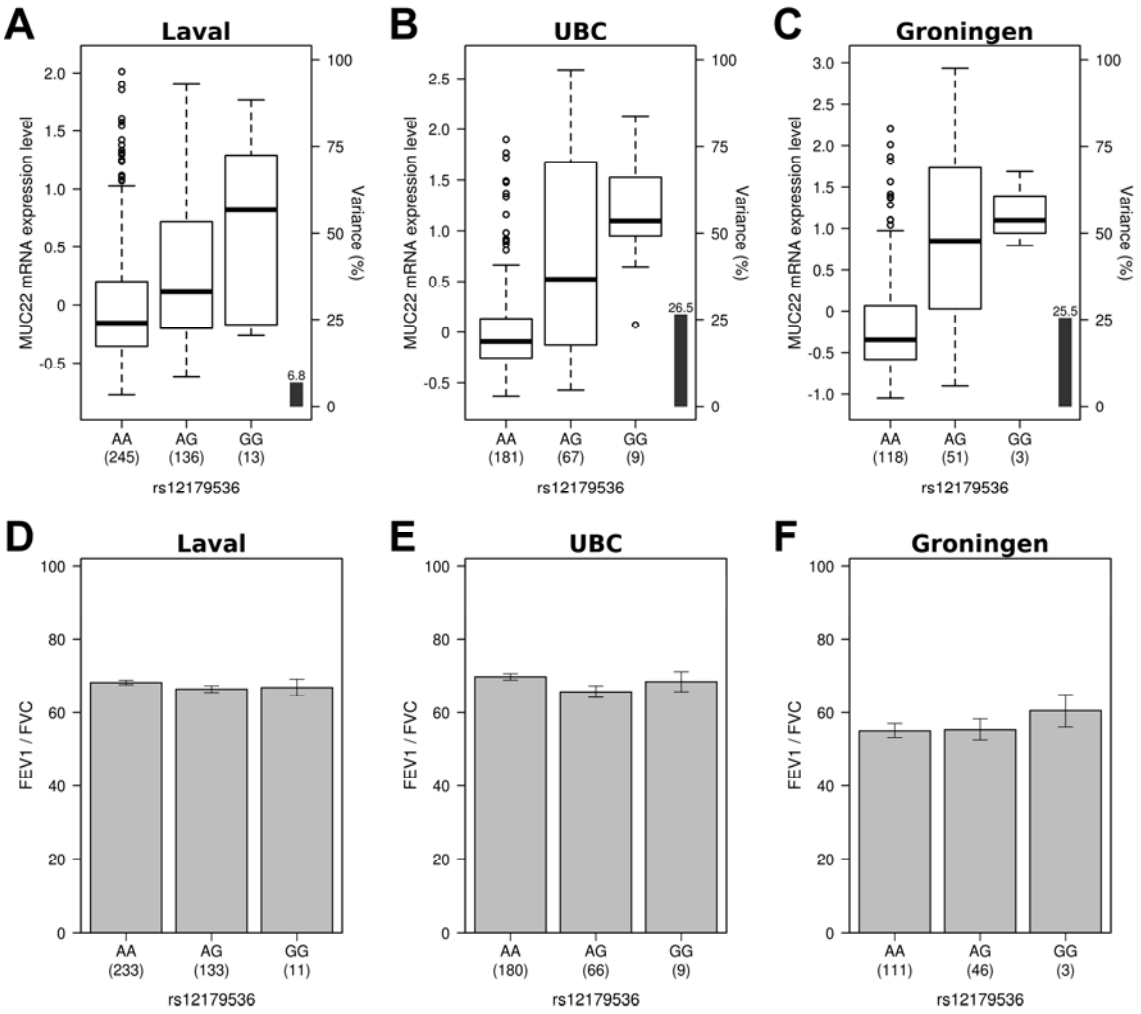


Figure S13. Direction of effects for causality model with rs2287765, mRNA expression of SPINK5, and FEV1/FVC. Data are presented as described in Figure 2.

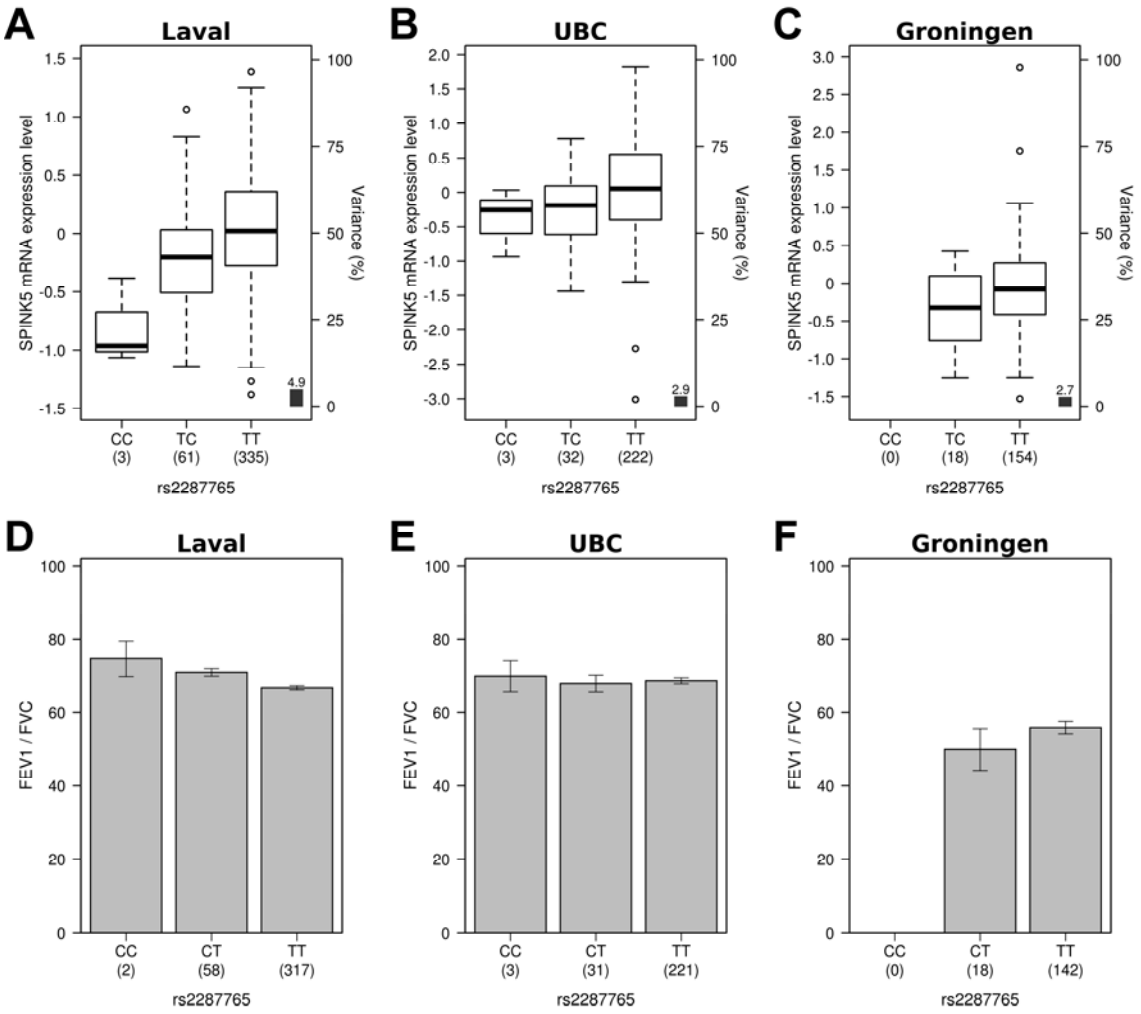


Figure S14. Causality genes in the Aryl Hydrocarbon Receptor Signaling pathway. Causality genes identified in this study are in red.

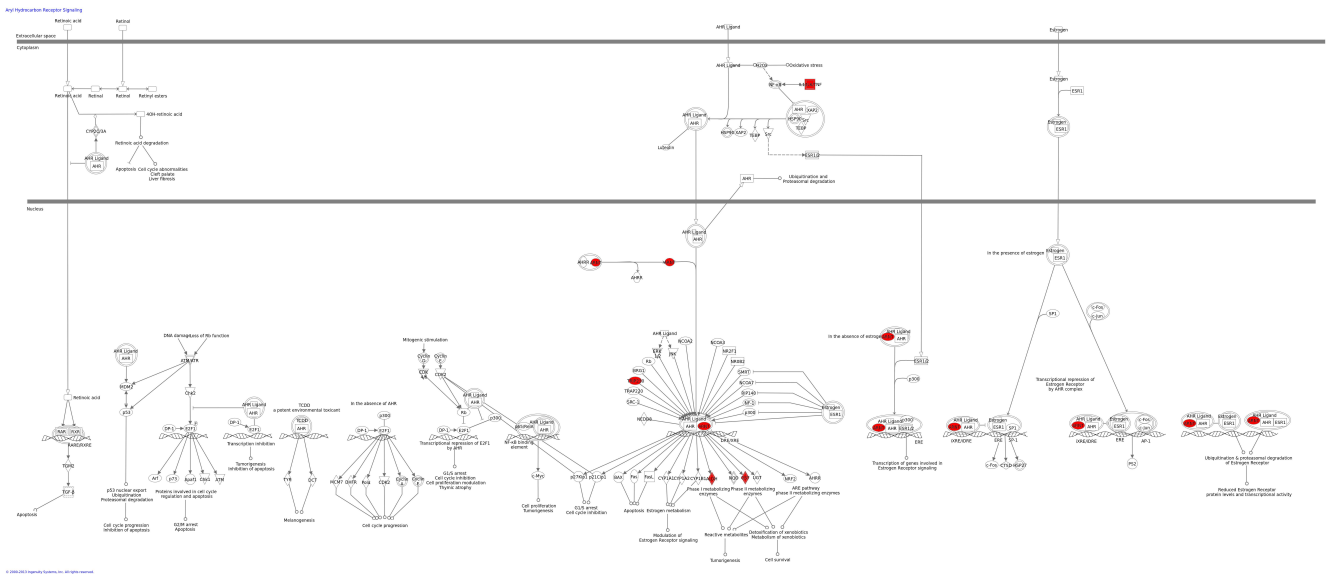


Figure S15. Causality genes in the Xenobiotic Metabolism Signaling pathway. Causality genes identified in this study are in red.

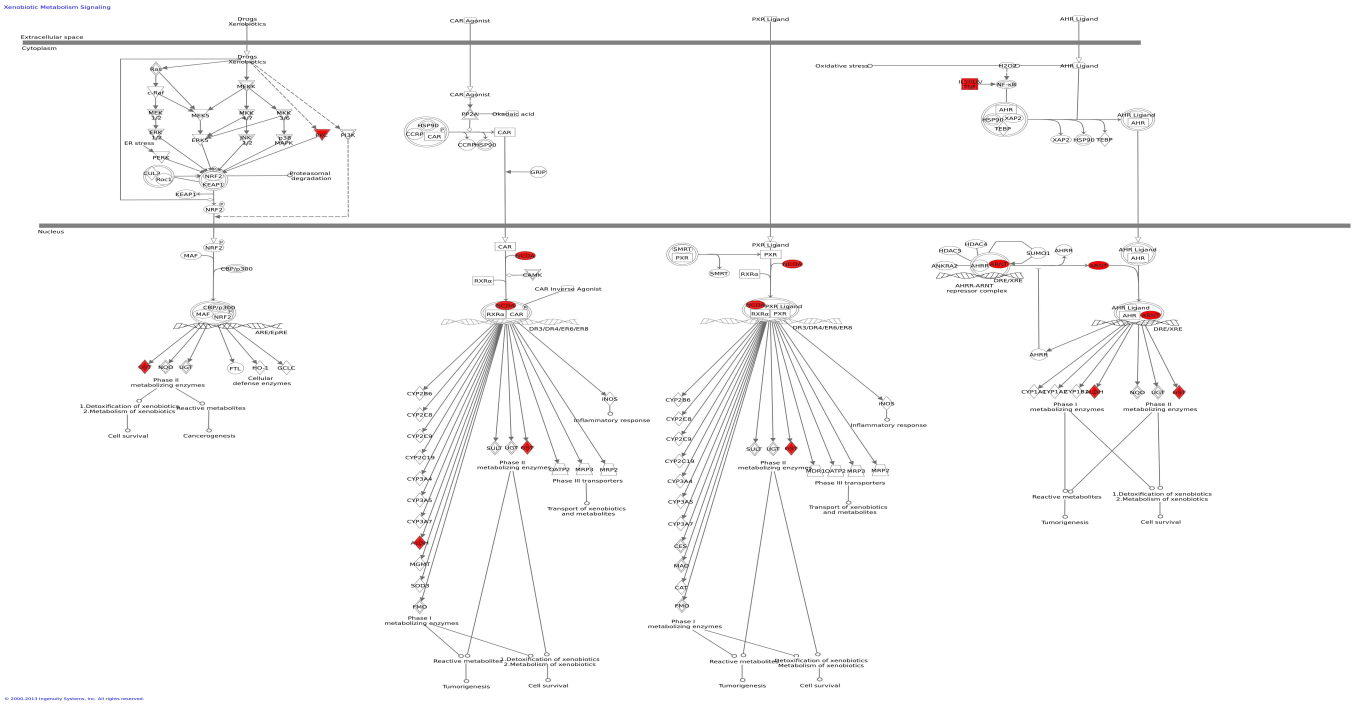


Figure S16. The glutathione metabolism pathway connecting the cyanoamino acid and the taurine/hypotaurine metabolism pathways.

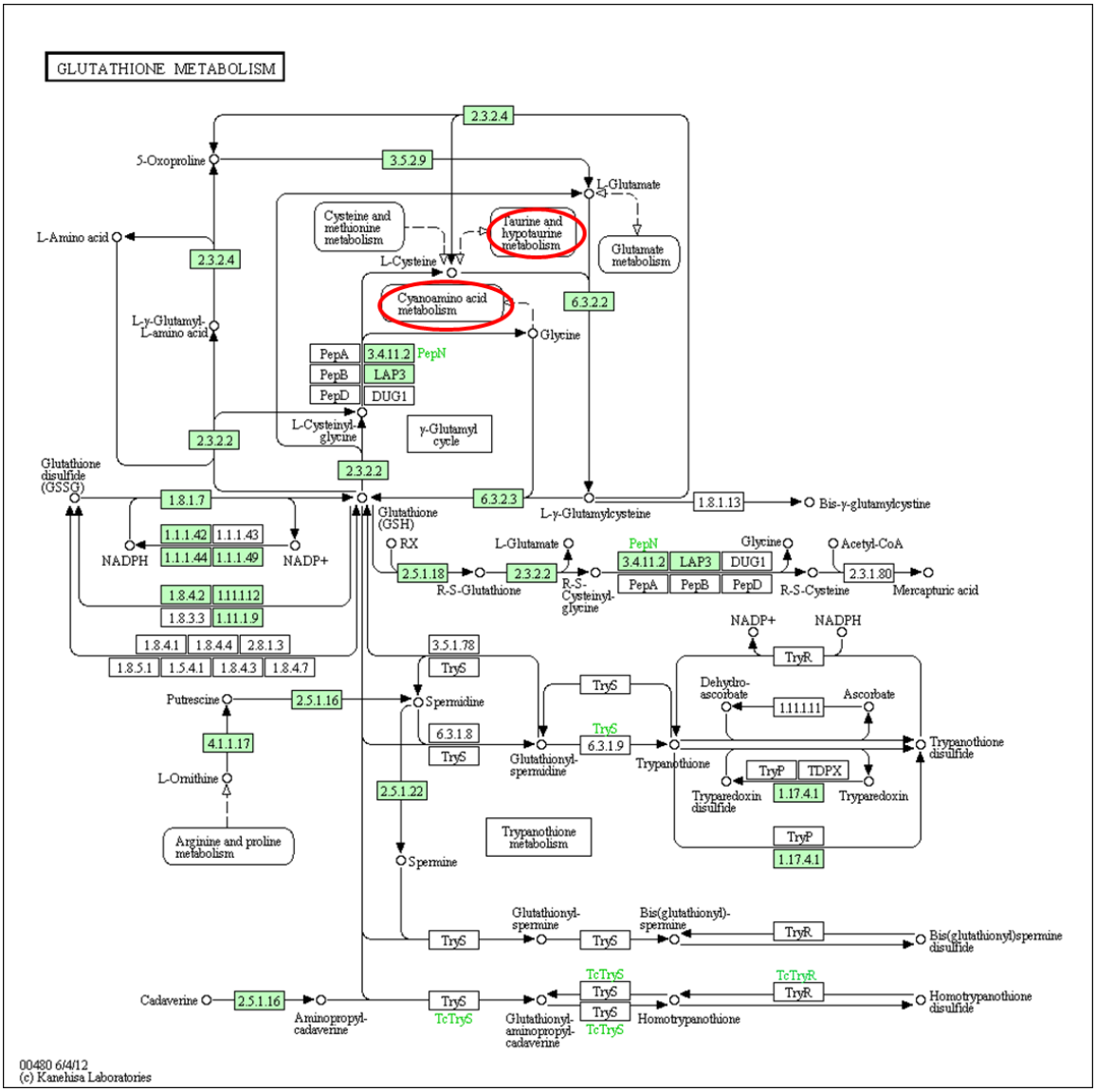
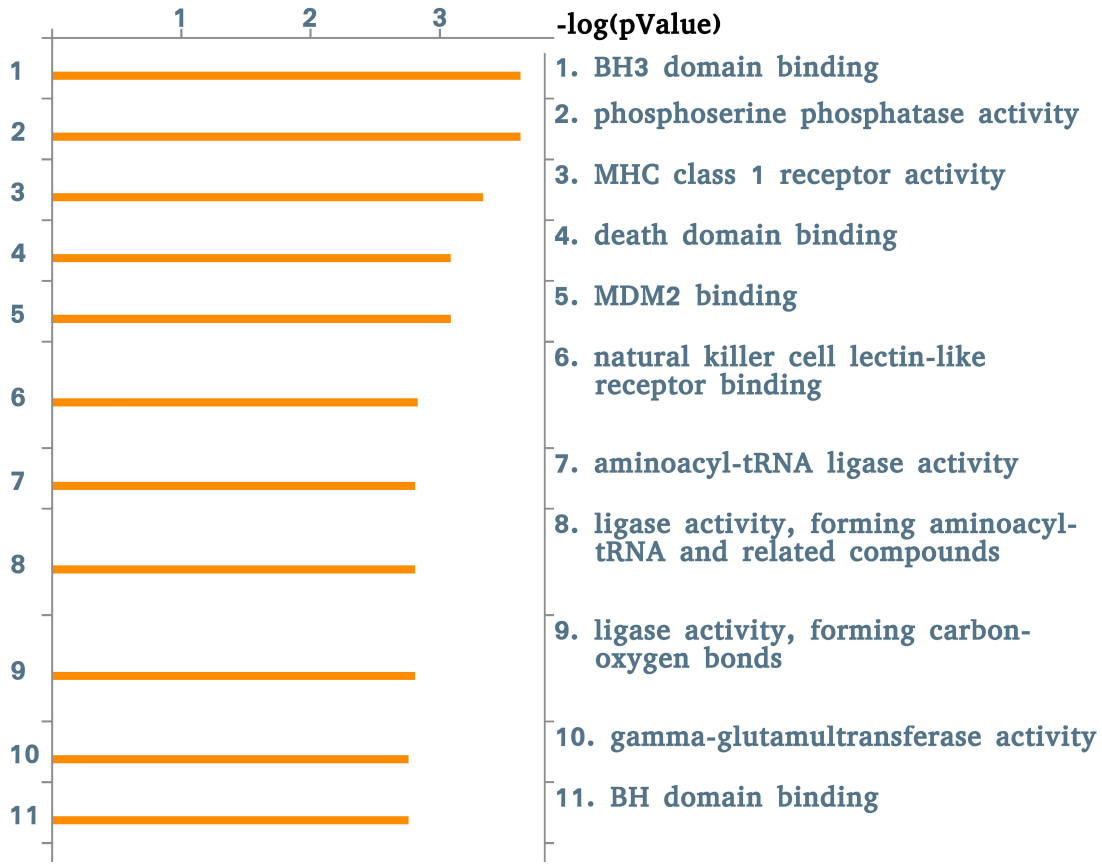


Figure S17. Gene Ontology (GO) project functional categories enriched for causality genes


1
2
3
-log(pValue)
1
2
3
4
5
6
7
8
9
10
11
1. BH3 domain binding
2. phosphoserine phosphatase activity
3. MHC class 1 receptor activity
4. death domain binding
5. MDM2 binding
6. natural killer cell lectin-like receptor binding
7. aminoacyl-tRNA ligase activity
8. ligase activity, forming aminoacyl-tRNA and related compounds
9. ligase activity, forming carbon-oxygen bonds
10. gamma-glutamultransferase activity
11. BH domain binding