A bioinformatics strategy for detecting the complexity of Chronic Obstructive Pulmonary Disease in Northern Chinese Han Population

Lin Hua^{1*†}, Li An^{2†}, Lin Li¹, Yongbiao Zhang³ and Chen Wang²

¹Biomedical Engineering Institute of Capital Medical University, Beijing, China

²Beijing Key Laboratory of Respiratory and Pulmonary Circulation Disorders, Beijing Institute of Respiratory Medicine, Department of Respiratory and Critical Care Medicine, Beijing Chao-Yang Hospital, Capital Medical University, Beijing, China

³CAS Key Laboratory of Genome Sciences and Information, Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing, China

(Received 1 August 2011, accepted 22 April 2012)

Chronic Obstructive Pulmonary Disease (COPD) is a complex human disease which is driven not only by genetic factors, but also by various environmental variables, such as gender, age and smoking. Therefore, there is a demand for investigating the complexity among various risk factors involved in COPD. In this study, 44 tagging SNPs from *EPHX1*, *GSTP1*, *SERPINE2* and *TGFB1* were selected and genotyped in 310 COPD cases and 203 controls, all of which belong to the Han from North China. We integrated functional prediction algorithms of nonsynonymous SNPs (nsSNPs) into Bayesian network to explore the complex regulatory relationships among disease traits and various risk factors. The results showed that three basic variables (age, sex and smoking) were risk factors of COPD-related trait and phenotype. Besides these environmental risk factors, deleterious nsSNPs were found to perform better than those of significant synonymous SNPs when used as variables to make risk prediction of disease outcome. This study provides further evidences for detecting the complexity of COPD in Northern Chinese Han Population.

Key words: Bayesian network, COPD, function score, nsSNP

INTRODUCTION

Chronic Obstructive Pulmonary Disease (COPD) is an inherently heterogeneous disorder. Within a given individual, there may be varying contributions of emphysema, chronic bronchitis, and long-term smoking. Although smoking is the important environmental risk factor, the existing reports show that only 10% of the chronic heavy smokers develop symptomatic COPD (Sethi and Rochester, 2000; Snider, 1989). Recently, a series of studies have implicated that COPD represents a complex disease with genetics contributions from multiple genes. For example, Vibhuti et al. (2007) approved that 113H/139H alleles of *mEPHX* and 105V/114V alleles of *GSTP1* and the combination of genotypes with same alleles are associated with imbalanced oxidative stress

and lung function in COPD patients. A genome-wide linkage analysis in the Boston Early-Onset COPD study suggested that *SERPINE2* may be associated with the COPD-related phenotypes (Palmer et al., 2003; Silverman et al., 2002). Similarly, the association analysis of the family-based data showed significant association of multiple *SERPINE2* single nucleotide polymorphisms (SNPs) with COPD related-phenotypes, which was confirmed by DeMeo et al. (2006) and Zhu et al. (2007) in two large European population-based association studies. Furthermore, it has been reported that polymorphisms in *TGFB1* seem to be associated with COPD-related traits in European population (Hersh et al., 2006). These findings provide a new insight to understand the etiology of COPD. By summarizing a group body of evidences, we found *EPHX1*, *GSTP1* and *SERPINE2* are reported to be associated with COPD in European (Brøgger et al., 2006; Penyige et al., 2010; Zhu et al., 2007) and east Asian such as Chinese (Xiao et al., 2003; Zhong et al., 2009) and Japanese population (Budhi et al., 2003). *TGFB1* shows more association in Chinese (Dai-shun et al., 2010; Ito et

Edited by Yoko Satta

* Corresponding author. E-mail: hualin7750@yahoo.com.cn

† These authors contributed to this work equally

Competing interests: The authors declare that they have no competing interests

al., 2008; Mak et al., 2009; Su et al., 2005) than in European (van Diemen et al., 2010; Wu et al., 2004). Given the large differences in genetic background of Asians and Europeans (Li et al., 2008; Zhao and Lee, 1989), it is important to provide further evidences related to these distinct populations to detect the complexity of COPD.

In recent years, genome-wide association studies (GWAS) have become an increasingly effective tool to identify genetic variation associated with the risk of complex diseases (Gauderman et al., 2007). However, currently identified genetic variants collectively can explain only a small proportion of disease phenotypic variance, and the noise causes many of identified signals are false positive loci. A noteworthy observation is that the major source of identification genetic variants associated with complex diseases is based on single base changes in the DNA sequence, some of which lead to the alterations in protein structure and function (Cavallo and Martin, 2005). It is therefore suggested that not all of the SNPs are equally functionally important. Generally, SNPs occurring in coding regions and causing an amino acid substitution or insertion of a stop codon are defined as nonsynonymous SNPs (nsSNPs). Amino acid substitutions might affect the formation of the protein in any step from DNA replication to post-translational modifications of the protein and could result in substantive changes in the protein structure and function. We can see that these nsSNPs are likely to affect the function of the proteins accounting for susceptibility to complex disease for their altering the encoded amino acid sequence. For example, Stenson et al. (2008) approved that among all of genetic changes involved in human diseases, half of which are attributed to nsSNP variants. Therefore, focus on those nsSNPs might help to identify the true susceptibility loci of complex diseases. In the present study, we genotyped 44 tagging SNPs spanning four genes (*EPHX1*, *GSTP1*, *SERPINE2* and *TGFB1*) as candidate genetic factors of COPD, and attempted to determine the association of them in Chinese Hans. To reflect the functional difference between nonsynonymous and synonymous SNPs, we performed the separated analysis for them.

Despite growing evidence for genetic importance of COPD-related traits, it is noteworthy that the findings about genetic contributions to COPD can only explain a small proportion of disease phenotypic variance because these disease traits are driven not only by genetic factors, but also by various environmental factors and their intricate interactions. It is therefore suggested that the challenge task is to characterize how the complex interactions among the environmental variables, genetic factors, and quantitative traits leading to the disease outcome. In the present study, we applied Bayesian network analysis to address this issue. As a promising tool, Bayesian network has been widely used to construct probabilistic

graphical models among different variables (Neil et al., 2005). Compared with other methods, Bayesian network has the advantage of uncovering conditional independency among variables, which provides a good way to survey direct interaction of variables (Neil et al., 2005). To explore the importance of nonsynonymous SNPs and synonymous SNPs in predicting COPD risk, we constructed two Bayesian networks using deleterious nonsynonymous SNPs predicted by three functional prediction algorithms and significant synonymous SNPs extracted by single association analysis, respectively. In addition, we used variables selected by Bayesian networks to perform risk prediction with logistic regression analysis. Our study provides a new insight into the genetic and environmental mechanism of COPD-related trait and phenotype.

MATERIALS AND METHODS

Study subjects We recruited 310 unrelated COPD patients aged 40–75 years from the respiratory outpatient clinics at 12 hospitals in Beijing from October of 2007 to March 2009 (An et al., 2011). The entry criteria were as follows: physician-diagnosed COPD; pulmonary function test showing post-bronchodilator forced expiratory volume in one second (FEV_1)/forced vital capacity (FVC) of less than 0.7 and FEV_1 of less than 0.8 predicted (The Global Strategy for the Diagnosis, Management and Prevention of COPD, Global Initiative for Chronic Obstructive Lung Disease (GOLD) 2006); and, no evidence of primary asthma or other respiratory diseases. Our control group comprised of 203 subjects with the same age range, which was recruited from the medical examination center at the same 12 hospitals during the same period as the case group. They have no respiratory symptoms history and exhibit normal pulmonary function of FEV_1 /FVC of more than 0.7 and FEV_1 of more than 0.8 predicted. Written informed consent was obtained from every participating subject, and the study protocol was approved by the research ethics boards of all participating hospitals.

Genotyping of SNPs Because linkage disequilibrium (LD) naturally exists in the human genome, a small number of tagging SNPs are sufficient to capture most of the genetic variation in high LD regions (Johnson et al., 2001). We selected our tagging SNPs according to the following three steps (An et al., 2012). Firstly, we listed the SNPs that previously showed significant relationship with COPD and its related phenotypes. Secondly, we directly downloaded the genotype data for the genes *EPHX1*, *GSTP1*, *SERPINE2* and *TGFB1* and their promoter regions from the public HapMap database using the Haploview software (Haploview, version 4.1; release 21; CHB+JPT panel). Then, we selected the tags according to their ability to tag LD blocks using Tagger program

from Haploview (Castaldi et al., 2009). Finally, 44 tagging SNPs ($MAF > 0.05$) were selected to capture the common variants of these four genes under pairwise mode with r^2 threshold of 0.8. Among these 44 SNPs, 8 SNPs are approved to be nsSNPs (Table 1). Genomic DNA was isolated from whole blood leukocytes by the conventional phenol-chloroform method. SNPs were genotyped using Illumina VeraCode technology performed on BeadXpress genotyping platform (Illumina Inc., USA). There was no significant departure from Hardy-Weinberg equilibrium (HWE) for all SNPs in control subjects ($p > 0.05$) by using a goodness-of-fit Chi-square test. An association analysis based on genotype Chi-square test was performed with PLINK software (<http://pngu.mgh.harvard.edu/~purcell/plink/>), and 8 SNPs ($p < 0.05$) displayed the significant difference between the COPD cases and controls (Table 1).

Prediction of the potential functional effect scores of nsSNPs In this progress, we applied three programs: VarioWatch (<http://genepipe.ngc.sinica.edu.tw/variowatch/>), SIFT (<http://sift.jcvi.org/>) and PolyPhen-2 (<http://genetics.bwh.harvard.edu/pph2/>) to estimate the risk effect of nsSNPs to COPD. The description of these three programs is as follows:

1) *VarioWatch*

This program incorporates 6 databases and can use different criteria to predict the function of the variant. The input is in the form of chromosome number and position, and it outputs predictions in terms of five risk levels including very low, low, medium, high and very high. In this analysis, we quantified these risk levels as five values: 0.2, 0.4, 0.6, 0.8 and 1 respectively. The first two risk levels were considered as tolerant and the last three levels were considered as deleterious (Chen et al., 2008).

2) *SIFT*

This program use a multi-step, sequence homology-base algorithm to predict tolerated and deleterious substitutions for nsSNPs. SIFT prediction is based on the evolutionary conservation of the amino acids within protein families. A highly conserved position is tending to be intolerant to substitutions, whereas a lowly conserved position should appear to tolerate most substitutions (Kumar et al., 2009). SIFT predicts an nsSNP to be ‘damaging’ (0.00–0.05) or to be ‘tolerated’ (0.05–1.00).

3) *PolyPhen-2*

Different from SIFT program which utilizes the protein structure information, PolyPhen-2 predicts functional effects of nsSNPs using eight sequence-based and three structure-based features (Adzhubei et al., 2010). Sequence-based features include whether the variant is in an active or binding site while structural components include whether the variant alters the polarity of the structure, potentially changing the hydrophobic core of a protein, or its interactions with itself or other

proteins. PolyPhen-2 calculates the naïve Bayes posterior probability of an nsSNP and classifies it into ‘benign’ (0–0.2), or ‘possibly damaging’ (0.2–0.85), or ‘probably damaging’ (0.85–1) groups. We can see that the higher PolyPhen-2 score indicates the greater ‘damaging’ effect. Note that the SIFT score shows the opposite trend in which the lower SIFT score implies the greater ‘damaging’ effect. In other words, the relationships between two functional effect scores (SIFT and PolyPhen-2) and their ‘damaging’ effects show an opposite trend. In order to keep the positive correlation of these two scores, we therefore used “1-SIFT” to express the functional effects of nsSNPs in the following result section.

In the present study, we considered those nsSNPs as deleterious nsSNPs when at least two programs predicted them as ‘deleterious’ or ‘damaging’.

Construction of Bayesian networks To further explore the roles of COPD-related environmental factors and genetic factors, we constructed Bayesian networks to dissect the complex regulatory relationships among these factors. A Bayesian network is a graphical model that encodes probabilistic relationships among variables of interest. Let $X = \{x_1, x_2, \dots, x_n\}$ be a set of variables. A Bayesian network over a set of variables is defined as a network structure S , which is a directed acyclic graph (DAG) over X and a set P of local probability distributions. The Bayesian network S encodes the assertions of conditional independence, i.e., each variable x_i is independent of its non-descendants, given its parents in S . In the present study, we allow the nodes fed into the Bayesian network contain the followings: Environmental variables (age, sex, and smoking), COPD phenotype and SNPs. Where, we classified those patients with > 5 cigarettes of pack-years into smoking group, and other patients into non-smoking group. On the other hand, we know that SNPs often regulate disease-related quantitative traits which lead to the manifestation of disease phenotypes. Therefore, we allow COPD-related quantitative trait (FEV1) come into Bayesian network. Specially, we also want to examine whether the roles of deleterious nsSNPs are more important than synonymous SNPs in the network, we therefore constructed four Bayesian networks in which the nodes were described as following: 1) BN1: Environmental variables (age, sex, and smoking) + COPD phenotype + deleterious nsSNPs; 2) BN2: Environmental variables (age, sex, and smoking) + COPD phenotype + significant SNPs ($p < 0.01$); 3) BN3: Environmental variables (age, sex, and smoking) + FEV1 + deleterious nsSNPs; 4) BN4: Environmental variables (age, sex, and smoking) + FEV1 + significant SNPs ($p < 0.01$). The conditional likelihood of the variables given their parents is represented in a Bayesian network by using Gaussian conditional densities. Under the assump-

Table 1. Forty-four captured tagging SNPs involved in four genes (*EPHX1*, *SERPINE2*, *GSTP1* and *TGFB1*) and their association with COPD using genotype-based Chi-square tests

Gene	SNP	Chromosome	Alleles	Region	Chi-square	<i>p</i> -value
<i>EPHX1</i>	rs1877724	1	C/T	intronic	0.1388	0.7095
<i>EPHX1</i>	rs1051740 [#]	1	C/T	coding	1.602	0.2056
<i>EPHX1</i>	rs1051741 [#]	1	C/T	coding	2.696	0.1006
<i>EPHX1</i>	rs2854450	1	C/T	5upstream	0.04641	0.8294
<i>EPHX1</i>	rs2292558	1	C/G	intronic	0.01017	0.9197
<i>EPHX1</i>	rs2260863	1	C/G	intronic	2.826	0.09273
<i>EPHX1</i>	rs868966	1	A/G	intronic	0.4563	0.4993
<i>EPHX1</i>	rs1009668 [#]	1	A/G	coding	0.2504	0.6168
<i>EPHX1</i>	rs41266229	1	A/G	intronic	16.19	5.73E-05**
<i>EPHX1</i>	rs2292568 [#]	1	C/T	coding	0.2384	0.6254
<i>EPHX1</i>	rs3766934	1	G/T	intronic	4.806	0.02837*
<i>EPHX1</i>	rs3738040	1	A/G	5upstream	0.6809	0.4093
<i>EPHX1</i>	rs2234922 [#]	1	A/G	coding	0.02807	0.8669
<i>SERPINE2</i>	rs6719480	2	C/T	intronic	1.283	0.2573
<i>SERPINE2</i>	rs4674841	2	G/T	intronic	1.445	0.2294
<i>SERPINE2</i>	rs17196253	2	A/G	intronic	2.569	0.1095
<i>SERPINE2</i>	rs975278	2	A/G	intronic	8.096	0.008054**
<i>SERPINE2</i>	rs920251	2	C/T	intronic	1.551	0.3050
<i>SERPINE2</i>	rs6748795	2	C/G	intronic	3.406	0.06495
<i>SERPINE2</i>	rs3820766	2	C/T	intronic	7.607	0.005814**
<i>SERPINE2</i>	rs10191694	2	A/C	intronic	2.278	0.2390
<i>SERPINE2</i>	rs7579646	2	A/G	intronic	2.232	0.1352
<i>SERPINE2</i>	rs13392495	2	A/G	intronic	4.518	0.03354*
<i>SERPINE2</i>	rs282254	2	C/T	intronic	0.002744	0.9582
<i>SERPINE2</i>	rs7583463	2	A/C	intronic	1.429	0.193
<i>SERPINE2</i>	rs729631	2	C/G	intronic	12.1	0.000867**
<i>SERPINE2</i>	rs6738983	2	C/T	intronic	0.03164	0.8588
<i>SERPINE2</i>	rs4674843	2	A/G	intronic	0.0715	0.3861
<i>SERPINE2</i>	rs861442	2	A/G	intronic	0.08717	0.5482
<i>SERPINE2</i>	rs7590948	2	A/G	intronic	1.384	0.3793
<i>SERPINE2</i>	rs2118409	2	C/G	intronic	0.952	0.6037
<i>SERPINE2</i>	rs6734100	2	C/G	intronic	6.516	0.01069*
<i>SERPINE2</i>	rs6712954 [#]	2	A/G	coding	1.721	0.2016
<i>SERPINE2</i>	rs6736436	2	C/T	3downstream	1.827	0.1765
<i>GSTP1</i>	rs4147581	11	C/G	intronic	0.2085	0.648
<i>GSTP1</i>	rs1138272 [#]	11	C/T	coding	7.971	0.00475**
<i>GSTP1</i>	rs1695 [#]	11	A/G	coding	0.2837	0.5943
<i>GSTP1</i>	rs947895	11	A/C	3downstream	0.003682	0.9516
<i>TGFB1</i>	rs6957	19	A/G	3utr	1.003	0.3165
<i>TGFB1</i>	rs2241715	19	G/T	intronic	1.076	0.2995
<i>TGFB1</i>	rs12980942	19	A/G	3downstream	0.05689	0.8115
<i>TGFB1</i>	rs2241718	19	C/T	3utr	0.0009921	0.9749
<i>TGFB1</i>	rs2241713	19	C/G	intronic	3.112	0.0777
<i>TGFB1</i>	rs1800469	19	C/T	3downstream	1.122	0.2894

Note: #nonsynonymous SNP (nsSNP); **p* < 0.05; ***p* < 0.01.

tion of parameter independence, an initial Bayesian network structure S is learned from the training data. From this initial network, greedy search algorithm with random restarts is performed to get the highest score posterior network to avoid local maxima. Finally, an optimized Bayesian network that maximizes the Bayesian factor is obtained using heuristic search of the network space in a specified domain. Note that some edges in the network might not be interpreted biologically; we exclude those edges that connect SNPs to environmental variables. The Bayesian networks were constructed using BNarray package of R software (<http://www.r-project.org>).

In addition, to validate our method, for those SNPs involved in each of Bayesian networks, we also used logistic regression models based on PLINK software to explore their epistasis interactions contributing to COPD phenotype. In this analysis, $p < 0.05$ indicates a significant epistasis interaction effect between two SNPs.

Prediction of COPD phenotype It is of great interest to use variables selected by Bayesian networks to perform risk prediction of COPD. In the present study, we applied four logistic regression models to predict the COPD risk; the variables included in the models are as following:

Model 1: Environment variables (age, sex, and smoking) + nsSNPs

Model 2: Environment variables (age, sex, and smoking) + deleterious nsSNPs extracted by Bayesian network (BN1)

Model 3: Environment variables (age, sex, and smoking) + significant SNPs ($p < 0.01$)

Model 4: Environment variables (age, sex, and smoking) + significant SNPs extracted by Bayesian network (BN2)

To compare the performance of prediction accuracy, we calculated the sensitivity, specificity, positive predictive value, negative predictive value and area under receiver operating characteristic (ROC) curve-AUC to examine the discrimination of different logistic regression models. Logically, logistic regression model including risk factors selected by Bayesian network should facilitate a better risk prediction than other models.

Comparison of classification performances To validate whether the nsSNPs altering the amino acids in important positions are more risk than general SNPs, we defined two SNP groups: one is the SNP group with 8 non-synonymous SNPs, and the other is the SNP group with 8 significant risk SNPs ($p < 0.05$). We applied four classifiers: naïve Bayes (John and Langley, 1995), k-Nearest Neighbor (kNN) (Gutin et al., 2002), Support Vector Machine (SVM) (Furey, 2000), and Random Forests (RF) (Pang et al., 2006) to compare the classification perfor-

mances of these two SNP groups when they were taken as predictor variables to classify samples. We used 5-fold cross validation to assess the classification accuracy rate of these different machine-learning methods. All samples are divided into five sets and in each analysis one set are considered as testing data, whereas the others are training data. We set k at three in k-Nearest Neighbor program and took Radial Basis Function (RBF) as the kernel function in the Support Vector Machine program. For Random Forests program, 5,000 trees were constructed.

RESULTS

Prediction of the potential functional effect scores of nsSNPs From 44 tagging SNPs, we found that rs729631 and rs975278 were significantly associated with COPD and the spirometry-related phenotype after Bonferroni correction. They were located within 2.79 kb of each other and were in strong linkage disequilibrium (LD) with an r^2 value of 0.978 in HapMap (CHB+JPT) data and were found to be in strong LD with an r^2 value of 0.848 in our study, and this region is approved a significant association with the risk of COPD by previous studies (DeMeo et al., 2006). In a family-based and population-based association study, Zhu and colleagues reported that COPD-related SNPs (rs16865421, rs6748795, rs975278, rs729631 and rs6734100) were also located in the same region (Zhu et al., 2007). Among 44 tagging SNPs genotyped, 8 nsSNPs were found to have potential functional effect scores predicted by VarioWatch, 1-SIFT or PolyPhen-2. As a result, five nsSNPs (rs1051740, rs1009668, rs1138272, rs1695 and rs6712954) are classified as deleterious by VarioWatch, two nsSNPs (rs1051740 and rs1138272) and one nsSNP (rs6712954) are predicted deleterious by 1-SIFT and PolyPhen-2, respectively (Table 2). It is also noted that a certain number of nsSNPs have very disagreement predicted scores using three different programs, which probably due to gene annotation errors or insufficient sequence evidence. The highest agreement was observed between 1-SIFT and PolyPhen-2 with a relative ratio agreement of 2/3, followed by the agreement between VarioWatch and 1-SIFT with a relative ratio agreement of 3/5, and the lowest agreement was seen between VarioWatch and PolyPhen-2 with a relative ratio agreement of 1/4. While previous studies have investigated the correlations between the 1-SIFT scores and PolyPhen-2 scores with Spearman's rank correlation coefficients, and significant correlations were found between these two scores (Zhu et al., 2008). According to be called 'deleterious' or 'damaging' by two programs at least, rs1051740 (*EPHX1*), rs1138272 (*GSTP1*) and rs6712954 (*SERPINE2*) are more likely to be potentially damaging than other variants which are likely benign. Interestingly, with Chi-square

Table 2. Prediction of the potential functional effect scores of nsSNPs

Gene	nsSNP	Chromosome	VarioWatch (effect)	1-SIFT (effect)	PolyPhen-2 (effect)
<i>EPHX1</i>	rs1051740	1	0.800 (deleterious)	0.990 (damaging)	NA
<i>EPHX1</i>	rs1009668	1	0.800 (deleterious)	0.270 (tolerant)	0.000 (benign)
<i>EPHX1</i>	rs2234922	1	0.400 (tolerant)	0.310 (tolerant)	NA
<i>EPHX1</i>	rs1051741	1	0.200 (tolerant)	NA	NA
<i>EPHX1</i>	rs2292568	1	0.400 (tolerant)	NA	NA
<i>GSTP1</i>	rs1138272	11	0.800 (deleterious)	0.950 (damaging)	0.018 (benign)
<i>GSTP1</i>	rs1695	11	0.800 (deleterious)	0.290 (tolerant)	0.000 (benign)
<i>SERPINE2</i>	rs6712954	2	0.800 (deleterious)	NA	0.999 (damaging)

Note: NA = not available due to the missing scores which can not be calculated by the corresponding program for this nsSNP.

tests based on genotypes, rs1051740 and rs6712954 were not found to be significant ($p = 0.2056$ and $p = 0.2016$ respectively) despite of their clearly damaging effects predicted.

In the present study, we incorporated three deleterious nsSNPs (rs1051740, rs1138272 and rs6712954) into the construction of Bayesian networks.

Construction of Bayesian networks In this analysis, we constructed Bayesian networks from four variables combinations as described in MATERIALS AND METHODS. While deleterious nsSNPs predicted by two programs at least are rs1051740, rs1138272 and rs6712954, and the genotype-based Chi-square tests show the five SNPs—rs41266229, rs975278, rs382076, rs729631 and rs1138272 are significantly different between the COPD cases and controls ($p < 0.01$). The Bayesian networks constructed from four nodes combinations named BN1, BN2, BN3 and BN4 were displayed in Fig. 1 (a–d), respectively, along with their corresponding probability tables. In each of Bayesian networks, an edge specified as node1→node2 indicates that node2 is a direct cause of node1.

From Fig. 1, we found that age, sex and smoking are substantial environmental risk factors of COPD phenotype and COPD-related quantitative trait (FEV1). This result agrees with the analysis performed by Kojima et al. (2005), in which they showed age and smoking were strong risk factors for COPD under the standard diagnostic criteria. By analyzing the genetic risk factors of COPD phenotype, it appeared that two deleterious nsSNPs (rs1138272 and rs6712954) in BN1 (Fig. 1a) and four significant SNPs (rs729631, rs3820766, rs1138272 and rs41266229) in BN2 (Fig. 1b) were associated directly with the risk of COPD phenotype. It is interesting to note that rs6712954 shows a clearly association with COPD phenotype in BN1 despite of its insignificant p -value ($p = 0.2016$) based on genotype test, and this association is approved by Kim et al. (2009). Moreover, from

BN2 we see rs975278 is a direct cause of rs729631. Indeed, rs729631 and rs975278 were found to be in strong LD with an r^2 value of 0.848, and this region was approved a significant association with the risk of COPD by previous studies. Although some interactions in the networks might be caused by LD, we can find other valuable interactions between SNPs from the constructed Bayesian networks. For example, from BN1, we found that rs1051740 (*EPHX1*) interacted with rs1138272 (*GSTP1*) or rs6712954 (*SERPINE2*). When we used PLINK software to explore their epistasis interactions, these interactions were also found to be significant: rs1051740 (*EPHX1*) $\times$ rs1138272 (*GSTP1*) ($p = 0.015$), and rs6712954 (*SERPINE2*) $\times$ rs1051740 (*EPHX1*) ($p = 0.035$). It has been reported that high *EPHX1* activity was associated with an increased risk for lifetime asthma which varied by *GSTP1* Ile105Val genotype (Salam et al., 2007). In addition, some SNPs showing more interactions with other SNPs in BN2 and BN4 can be explained by LD on one hand, and on the other hand, these interactions approved by epistasis interactions might be potential interactions between genes, such as rs975278 (*SERPINE2*) $\times$ rs41266229 (*EPHX1*) ($p = 0.011$). Although there are still some interactions which can't be approved by available evidences, it has been suggested that our method has great potential to detect novel gene-gene or gene-environment interactions affecting COPD susceptibility, severity, and response to treatment.

It is noted that only one nsSNP (rs1138272) in BN3 (Fig. 1c) and four significant SNPs (rs3820766, rs975278, rs729631 and rs1138272) in BN4 (Fig. 1d) connect the COPD-related trait (FEV1) directly. This result was approved by a linear regression analysis for FEV1 with age, sex and pack years of smoking as covariates. We found that rs1138272, rs729631, rs975278 and rs3820766 also exhibited a significant association after Bonferroni correction. While rs975278 was associated with the lower FEV1 ($\beta = -0.281$, $p_{Bon} = 0.015$), and rs729631 also presented strong association with the lower FEV1 ($\beta = -0.205$,

$p_{Bon} = 0.007$) (An et al., 2012). Note that rs1138272 is a genetic risk factor not only for COPD phenotype but also for quantitative trait FEV1, which is supported by the interactions between exposure to air pollution and *GSTP1* (rs1138272) for development of childhood allergic disease (Melén et al., 2008). Furthermore, although BN1 and BN2 are associated with COPD phenotype whereas BN3 and BN4 are associated with FEV1 trait (in this sense

BN1 and BN2 are different from BN3 and BN4), it is worth noting that one nsSNP (rs1138272) and three significant SNPs (rs3820766, rs1138272 and rs729631) were shared by Bayesian networks constructed using COPD phenotype and FEV1 trait. On one hand, this result can be explained by the strong association between COPD phenotype and FEV1 trait (Hersh et al., 2006); and on the other hand, recent advances in genetic studies found that

Fig. 1. Continued

Fig. 1. Four Bayesian networks constructed with different nodes combinations, along with their corresponding probability tables: a) BN1: environmental variables (age, sex, and smoking) + COPD phenotype + deleterious nsSNPs; b) BN2: environmental variables (age, sex, and smoking) + COPD phenotype + significant SNPs ($p < 0.01$); c) BN3: environmental variables (age, sex, and smoking) +FEV1+ deleterious nsSNPs; d) BN4: environmental variables (age, sex, and smoking) +FEV1+ significant SNPs ($p < 0.01$). In each of Bayesian networks, an edge specified as A→B indicates that B is a direct cause of A. For example, the direction of arrow from COPD to smoking (COPD→smoking) presented in BN1 indicates that smoking is the cause of COPD.

some SNPs, such as rs1138272 in *GSTP1*, might be important SNPs in the development of COPD (van Diemen et al., 2010).

To validate our results, we also performed two additional Bayesian network analyses, which are i) using low LD (independent) SNPs as nodes (Supplementary Fig.

Table 3. COPD risk prediction with four Logistic regression models

Classifiers	The Prediction Property (95% Confidence Interval)				
	Sensitivity (%)	Specificity (%)	Positive predictive value (%)	Negative predictive value (%)	Area under ROC curve (AUC) (%)
Model 1 (EV+8nsSNPs)	75.9 (70.6–80.7)	64.2 (56.5–71.3)	78.1 (72.8–82.8)	61.3 (53.8–68.5)	77.2 (72.9–81.5)
Model 2 (EV+2deleterious nsSNPs extracted by BN1)	80.5 (75.5–84.8)	75.3 (67.9–81.7)	85.9 (81.3–89.7)	67.4 (60.1–74.2)	85.3 (81.8–88.8)
Model 3 (EV+5significant SNPs ($p < 0.01$))	75.1 (69.8–79.9)	65.0 (57.2–72.3)	79.9 (74.7–84.4)	58.6 (51.0–65.8)	76.0 (72.6–80.5)
Model 4 (EV+4significant SNPs extracted by BN2)	74.0 (68.7–78.9)	63.8 (55.8–71.2)	79.5 (74.3–84.1)	56.4 (48.8–63.7)	76.5 (72.2–80.9)

Note: EV: Environmental Variable (age, sex, and smoking).

Fig. 2. ROC curves obtained using four logistic regression models for detecting COPD, which are colored with different lines respectively.

S1); and ii) using deleterious nsSNPs, significant SNPs ($p < 0.01$), and low LD (independent) SNPs as nodes (Supplementary Fig. S2). To decrease network computation complexity, and assure the enough independent of SNPs, we used 0.3 of r^2 (LD value) as cutoff to filter SNPs. Eight low LD (independent) SNPs were extracted, they were: rs1877724, rs2234922, rs2854450, rs6734100, rs2118409, rs1009668, rs868966 and rs2260863. From Supplementary Figs. S1 and S2, we can find that among 8 low LD SNPs, only 1 nsSNP (rs2234922) and 1 significant SNP (rs6734100) are associated with COPD phenotype and FEV1 trait synchronously, which approves our results that deleterious and significant SNPs are more important and sensitive than

Fig. 3. Comparison of classification correct rate of two SNP groups using four classifiers. The two SNP groups are SNP group with 8 nonsynonymous SNPs (green cuboids) and SNP group with 8 significant risk SNPs ($p < 0.05$) (gray cuboids). Four classifiers are as follows: Random Forests (RF), Support Vector Machine (SVM), k-Nearest Neighbor (kNN) and naïve Bayes.

other common SNPs in contributing to disease. In other words, significant SNPs and nonsynonymy SNPs are all worthy focused on in genetics association studies.

Prediction of COPD phenotype using variables selected by Bayesian networks In this study, we applied four logistic regression models as described in MATERIALS AND METHODS to predict the COPD risk. We calculated the sensitivity, specificity, positive predictive value, negative predictive value and area under ROC curve (AUC) of four logistic regression models for detecting COPD (Table 3). As expected, logistic regression models including risk factors selected by Bayesian networks performed a better risk prediction than other models (Fig. 2). For example, the AUC score

Table 4. Sensitivity, specificity, positive predictive value, negative predictive value, and area under ROC curve (AUC) for SNP group with 8 nsSNPs used to detect COPD in four classifiers

Classifiers	The Prediction Property (95% Confidence Interval)				
	Sensitivity (%)	Specificity (%)	Positive predictive value (%)	Negative predictive value (%)	Area under ROC curve (AUC) (%)
Random Forests	97.9 (95.5–99.2)	69.1 (61.8–75.7)	83.2 (78.8–87.1)	95.4 (90.3–98.3)	89.3 (86.1–92.5)
SVM	95.8 (92.7–97.8)	66.9 (59.5–73.7)	81.9 (77.4–85.9)	91.0 (84.8–95.3)	86.5 (82.7–90.2)
kNN	91.9 (88.1–94.8)	63.5 (56.1–70.1)	79.8 (75.1–84.0)	83.3 (76.1–89.1)	81.6 (77.2–86.0)
naïve Bayes	87.0 (82.5–90.7)	66.3 (58.9–73.1)	80.2 (75.3–84.5)	76.4 (69.0–82.8)	78.3 (73.7–82.9)

increased 8.1% from model 1 to model 2, and 0.5% from model 3 to model 4. Note that model 2 shows the best prediction accuracy (AUC score: 85.3%) whereas the remaining three models only show a slight difference in prediction, which suggests that environmental variables and deleterious nsSNPs are more sensitive and specific in predicting disease, and Bayesian network can further this power.

Comparison of classification performances The classification results are completely consistent with what we expected, and four classifiers all showed that the SNP group with 8 nonsynonymous SNPs was more powerful than the other group when used as predictor variables to classify samples (Fig. 3). This result supports our hypothesis and can indicate that some nsSNPs altering the amino acids in important positions and resulting in functional change of the corresponding proteins might contribute to disease phenotype. The sensitivity, specificity, positive predictive value, negative predictive value, and area under ROC curve (AUC) for two SNP groups used to detect COPD in four classifiers were shown in Table 4.

Note that for both of SNP groups, Random Forests classifier performed better (86.7% for SNP group with 8 nsSNPs, compared with 69.0% for SNP group with 8 significant SNPs) than other three classifiers. Specially, it is often interest to know which of SNPs are important in classification in Random Forests program. There are two measures of importance in Random Forests, the mean decrease in accuracy and the mean decrease Gini index (MDG) (Pang et al., 2006). In this analysis, we used MDG to measure the risk of a SNP. Greater MDG will indicate that the degree of impurity arising from category could be reduced farthest by a SNP, and thus suggests an important SNP. We ranked 8 nsSNPs according to their MDG (Supplementary Table S1), and found that rs6712954 ranked first with the largest MDG (MDG = 0.7824), followed by rs1138272 (MDG = 0.6932). Interestingly, this result agrees with the constructed Bayesian network (BN1), in which rs1138272 and rs6712954 are successfully identified as two deleterious nsSNPs associated with the risk of COPD phenotype.

DISCUSSION AND CONCLUSION

In this study, we provided the different genotyping data of the relationship between four genes (*EPHX1*, *SERPINE2*, *GSTP1*, and *TGFB1*) and COPD phenotype in Chinese Han population. Considering those currently identified genetic variants can explain only a very small proportion of disease phenotypic variance, it has become a popular belief that complex disease traits are driven not only by significant SNPs, but also by functional SNPs. On the other hand, nonsynonymous SNPs can produce a different peptide sequence; they are more likely to be disease causal variants than synonymous SNPs. As a result of causal SNPs' small effect size (mean OR < 1.4 for most common human diseases) and the multiple testing burden, the results of many association studies including a mass of SNPs are actually false positives (Kang et al., 2011), which causes the difficult replication of genetic association studies. Some biology evidences indicate that SNPs do not directly cause diseases; instead, they often regulate disease-related quantitative traits such as gene expression, which in term lead to the manifestation of downstream disease phenotypes. We therefore incorporated three functional prediction algorithms of nsSNPs into Bayesian network to explore the complex regulatory relationships among disease traits and various risk factors. In fact, if a SNP is highly associated with disease phenotype and related to functional information, this SNP is more likely to be a true disease association signal. Our results support that three basic variables (age, sex and smoking) are substantial risk factors of COPD-related trait (FEV1) and phenotype. Moreover, deleterious nsSNPs were found to perform better than those of significant synonymous SNPs when used as variables to make risk prediction of disease outcome. We compared our study with others and noted that few of the previous studies mentioned the discrimination of nsSNPs and synonymous SNPs in the association with COPD disease. Despite of the inevitable prediction errors, predicting disease phenotype using our methods might provide a good way to explore the relationship between risk factors and susceptibility to COPD.

Using Bayesian network, it is encouraging to observe the association of the *SERPINE2* (rs6712954, rs729631,

rs3820766 and rs975278) and *GSTP1* (rs1138272) with the COPD in Chinese Han population. In fact, for genes *SERPINE2* and *GSTP1*, there are many previous reports approving their association with COPD. *SERPINE2* has been shown to inhibit several trypsin-like serine proteases, including thrombin, urokinase, and plasmin, which play important roles in inflammation and wound repair following tissue injury (Baker et al., 1980; Scott et al., 1985). Moreover, it has been reported that the expression of *SERPINE2* was up-regulated by interleukin-1 β , tumor necrosis factor- α and transforming growth factor- β , which have been suggested to be involved in the development of COPD (Mbebi et al., 1999; Vaughan and Cunningham, 1993). Although the physical function of *SERPINE2* in COPD and emphysema has not yet been fully resolved, it is suggested that *SERPINE2* may play an important role in the pathogenesis of COPD. For another COPD-related gene *GSTP1*, it has been reported that 105V/114V alleles of *GSTP1* are associated with imbalanced oxidative stress and lung function in patients (Vibhuti et al., 2007). Similarly, DeMeo et al. (2007) concluded that the apical-predominant emphysematous destruction is influenced by polymorphisms in the xenophobic enzymes of *GSTP1*. In summary, Han Chinese shares the same COPD susceptibility genes of *SERPINE2* and *GSTP1* with Europeans. The association of these two genes with COPD needs to be validated in more ethnic groups and in larger populations.

Furthermore, by comparing the Bayesian network topologies obtained from the deleterious nsSNPs and significant synonymous SNPs (besides rs1138272) respectively, it appears that COPD-related trait and phenotype are driven by these different types of SNPs. Note that nsSNPs are more closely associated with COPD phenotype, suggesting that it is important to consider those functional loci altering the amino acids in important positions involved in those COPD-related genes even if they were insignificant in single association analysis. Therefore, our approaches are useful to understand the COPD mechanism in the case of COPD susceptibility genes even for a relative small sample. For example, one nsSNP downstream of *EPHX1* (rs1009668) resulting in replacement of valine with methionine at amino acid position 622, was approved to be significantly associated with the change in FEV1. This SNP which is associated with the increased percent emphysema and the reduced bronchodilator responsiveness (BDR) among COPD patients may have a role in COPD pathogenesis (Kim et al., 2009). Future investigation is to develop more models that incorporate different functional prediction scores into genetic and environmental factors and make the risk identification become more robust.

It is pointed out that the three programs (VarioWatch, SIFT and PolyPhen-2) used to predict the function of nsSNPs on protein structure and activity in our analysis

are main representatives in silico approaches. Although it has been reported that the false positive error in SIFT and PolyPhen-2 are only 19% and 9%, respectively (Ng and Henikoff, 2006), the prediction accuracy of silico algorithms was highly affected by interference of redundant motifs and the accuracy of phenotype information. Therefore, the phenotype of nsSNPs predicted as deleterious or damaging effect needs to be clarified in future studies by computation and experiment. In addition, we have to point out the limitation of our study in which SNPs used in this analysis can not reflect the causative alleles completely for some uncontrolled reasons. However, our study can help researchers to find the associated alleles (genes) to COPD, or these alleles (genes) might be potential signals in genome-wide association analysis. Moreover, the aim of our study is to find the complex interaction relationship between functional SNPs and environmental factors using Bayesian network, which can help detecting the potential interactions of multiple polymorphisms or environment factors thus may be important to understand the biology and biochemical processes of the disease etiology. In the future, we will consider extend the number of SNPs to perform our study, including a genome-wide association study, if possible.

It is noteworthy that the sample size of our study was relatively small, which might cause the insufficient efficacy of our discoveries. We performed the power calculation using GeneticsDesign package of R software with the following conditions: frequency of A (protective) allele, 0.15; significance level, 0.05; prevalence of COPD, 10% according to the epidemiological study carried by Mannino (2002); and genetic model, additive. For the given sample size, the statistical power was nearly 80% (78.5%). In response to the relative small sample size, we incorporated the functional effect scores of nsSNPs and Bayesian network into genetic analysis to improve statistical power. It is encouraging to observe some significant results. Besides, note that we performed genotype-based Chi-square test in association analysis to keep consistent with our further analysis, including Bayesian network construction, logistic regression and classification analysis. In fact, the allele-based test is more sensitive than genotype-based Chi-square test in identifying significant SNPs. In the present study, the significant SNPs identified by allele-based test and genotype-based test respectively are completely consistent, and the correlation coefficient between p -values acquired with two kinds of tests is up to 0.871 ($p < 0.001$). We believe that the allele-based test will be helpful when applying association tests to a great number of SNPs, which will be considered in our future study to confirm these discoveries.

In conclusion, our results contribute to the knowledge base of risk factors to COPD-related trait and phenotype, particularly in the East Asian population. Certainly, the genetic heterogeneity between the study populations

could have contributed to the different results of the different population-based association studies. Further researches are warranted to confirm the observation and identify the exact environmental and genetic risk factors involved in COPD pathogenesis.

This work is supported by the National Natural Science Foundation of China (Grant Nos. 31100905) and the Science Technology Development Project of Beijing Municipal Commission of Education (SQKM201210025008). This study is also funded by the excellent talent cultivation project of Beijing, and supported by the foundation-clinical cooperation project of capital medical university (11JL33), the Beijing science and technology project (Z090507006209018) and the New Star Program of Beijing Science and Technology (2009A11).

REFERENCES

- Adzhubei, I. A., Schmidt, S., Peshkin, L., Ramensky, V. E., Gerasimova, A., Bork, P., Kondrashov, A. S., and Sunyaev, S. (2010) A method and server for predicting damaging missense mutations. *Nat. Methods* **7**, 248–249.
- An, L., Yang, T., Zhang, Y., Lin, Y., Zhang, H., Jiao, X., Hua, L., Dai, H., and Wang, C. (2012) Association of SERPINE2 gene with the risk of chronic obstructive pulmonary disease and spirometric phenotypes in northern Han Chinese population. *Mol. Biol. Rep.* **39**, 1427–1433.
- Baker, J. B., Low, D. A., Simmer, R. L., and Cunningham, D. D. (1980) Protease-nexin: A cellular component that links thrombin and plasminogen activator and mediates their binding to cells. *Cell* **21**, 37–45.
- Brøgger, J., Steen, V. M., Eiken, H. G., Gulsvik, A., and Bakke, P. (2006) Genetic association between COPD and polymorphisms in TNF, ADRB2 and EPHX1. *Eur. Respir. J.* **27**, 682–688.
- Budhi, A., Hiyama, K., Isobe, T., Oshima, Y., Hara, H., Maeda, H., and Kohno, N. (2003) Genetic susceptibility for emphysematous changes of the lung in Japanese. *Int. J. Mol. Med.* **11**, 321–329.
- Castaldi, P. J., Hersh, C. P., Reilly, J. J., and Silverman, E. K. (2009) Genetic associations with hypoxemia and pulmonary arterial pressure in COPD. *Chest* **135**, 737–744.
- Cavallo, A., and Martin, A. C. R. (2005) Mapping SNPs to protein sequence and structure data. *Bioinformatics* **21**, 1443–1450.
- Chen, Y.-H., Liu, C.-K., Chang, S.-C., Lin, Y.-J., Tsai, M.-F., Chen, Y.-T., and Adam, Y. (2008) GenoWatch: a disease gene mining browser for association study. *Nucleic Acids Res.* **36** (suppl. 2), W336–W340.
- Dai-shun, L., Xiao-ou, L., Bin-wu, Y., Lei, C., Tao, W., Dan, X., and Fu-qiang, W. (2010) Effects of single nucleotide polymorphisms 869 T/C and 915 G/C in the exon 1 locus of transforming growth factor-beta1 gene on chronic obstructive pulmonary disease susceptibility in Chinese. *Chin. Med. J. (Engl)*, **123**, 390–394.
- DeMeo, D. L., Mariani, T. J., Lange, C., Srisuma, S., Litonjua, A. A., Cledón, J. C., Lake, S. L., Reilly, J. J., Chapman, H. A., Mecham, B. H., et al. (2006) The *SERPINE2* gene is associated with chronic obstructive pulmonary disease. *Am. J. Hum. Genet.* **78**, 253–264.
- DeMeo, D., Hersh, C., Hoffman, E., Litonjua, A., Lazarus, R., Sparrow, D., Benditt, J., Criner, G., Make, B., Martinez, F., et al. (2007) Genetic determinants of emphysema distribution in the national emphysema treatment trial. *Am. J. Respir. Crit. Care Med.* **176**, 42–48.
- Furey, T. S. (2000) Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics* **16**, 906–914.
- Gauderman, W. J., Murcray, C., Gilliland, F., and Conti, D. V. (2007) Testing association between disease and multiple SNPs in a candidate gene. *Genet. Epidemiol.* **31**, 383–395.
- Gutin, G., Yeo, A., and Zverovich, A. (2002) Traveling salesman should not be greedy: domination analysis of greedy-type heuristics for the TSP. *Discrete Appl. Math.* **117**, 81–86.
- Hersh, C. P., DeMeo, D. L., Lazarus, R., Cledón, J. C., Raby, B. A., Benditt, J. O., Criner, G., Make, B., Martinez, F. J., Scanlon, P. D., et al. (2006) Genetic association analysis of functional impairment in Chronic Obstructive Pulmonary Disease. *Am. J. Respir. Crit. Care Med.* **173**, 977–984.
- Ito, M., Hanaoka, M., Droma, Y., Hatayama, O., Sato, E., Katsuyama, Y., Fujimoto, K., and Ota, M. (2008) The association of transforming growth factor beta 1 gene polymorphisms with the emphysema phenotype of COPD in Japanese. *Intern. Med.* **47**, 1387–1394.
- John, G. H., and Langley, P. (1995) Estimating continuous distributions in Bayesian classifiers. *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, Morgan Kaufmann, San Mateo., 338–345.
- Johnson, G. C., Esposito, L., Barratt, B. J., Smith, A. N., Heward, J., Di Genova, G., Ueda, H., Cordell, H. J., Eaves, I. A., Dudbridge, F., et al. (2001) Haplotype tagging for the identification of common disease genes. *Nat. Genet.* **29**, 233–237.
- Kang, J., Zheng, W., Li, L., Lee, J. S., Yan, X., and Zhao, H. (2011) Use of Bayesian networks to dissect the complexity of genetic disease: application to the Genetic Analysis Workshop 17 simulated data. *BMC Proceedings* **5** (Suppl. 9), S37.
- Kim, W. J., Hersh, C. P., DeMeo, D. L., Reilly, J. J., and Silverman, E. K. (2009) Genetic association analysis of COPD candidate genes with bronchodilator responsiveness. *Respir. Med.* **103**, 552–557.
- Kojima, S., Sakakibara, H., Motani, S., Hirose, K., Mizuno, F., Ito, M., and Hashimoto, S. (2005) Effects of smoking and age on chronic obstructive pulmonary disease in Japan. *Journal of Epidemiology* **15**, 113–117.
- Kumar, P., Henikoff, S., and Ng, P. C. (2009) Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nature Protocols* **4**, 1073–1081.
- Li, J. Z., Absher, D. M., Tang, H., Southwick, A. M., Casto, A. M., Ramachandran, S., Cann, H. M., Barsh, G. S., Feldman, M., Cavalli-Sforza, L. L., et al. (2008) Worldwide human relationships inferred from genome-wide patterns of variation. *Science* **319**, 1100–1104.
- Mak, J. C. W., Chan-Yeung, M. M. W., Ho, S. P., Chan, K. S., Choo, K., Yee, K. S., Chau, C. H., Cheung, A. H. K., and Ip, M. S. M. (2009) Elevated plasma TGF-beta1 levels in patients with chronic obstructive pulmonary disease. *Respir. Med.* **103**, 1083–1089.
- Mannino, D. (2002) COPD: epidemiology, prevalence, morbidity, and mortality and, disease heterogeneity. *Chest* **121**, 121–126.
- Mebis, C., Hantaï, D., Jandrot-Perrus, M., Doyennette, M.-A., and Verdière-Sahuqué, M. (1999) Protease nexin I expression is up-regulated in human skeletal muscle by injury-related factors. *J. Cell. Physiol.* **179**, 305–314.
- Melén, E., Nyberg, F., Lindgren, C. M., Berglund, N., Zucchelli,

- M., Nordling, E., Hallberg, J., Svartengren, M., Morgenstern, R., Kere, J., et al. (2008) Interactions between glutathione S-Transferase P1, Tumor Necrosis Factor, and traffic-related air pollution for development of childhood allergic disease. *Environ. Health Perspect.* **116**, 1077–1084.
- Neil, M., Fenton, N., and Tailor, M. (2005) Using Bayesian networks to model expected and unexpected operational losses. *Risk Analysis*, **25**, 963–972.
- Ng, P. C., and Henikoff, S. (2006) Predicting the effects of amino acid substitutions on protein function. *Annu. Rev. Genomics Hum. Genet.* **7**, 61–80.
- Palmer, L. J., Celedón, J. C., Chapman, H. A., Speizer, F. E., Weiss, S. T., and Silverman, E. K. (2003) Genome-wide linkage analysis of bronchodilator responsiveness and post-bronchodilator spirometric phenotypes in chronic obstructive pulmonary disease. *Hum. Mol. Genet.* **12**, 1199–1210.
- Pang, H., Lin, A., Holford, M., Enerson, B. E., Lu, B., Lawton, M. P., Floyd, E., and Zhao, H. (2006) Pathway analysis using random forests classification and regression. *Bioinformatics* **22**, 2028–2036.
- Penyige, A., Poliska, S., Csanky, E., Scholtz, B., Dezso, B., Schmelcz, I., Kilty, I., Takacs, L., and Nagy, L. (2010) Analyses of association between PPAR gamma and *EPHX1* polymorphisms and susceptibility to COPD in a Hungarian cohort, a case-control study. *BMC Med. Genet.* **11**, 152–160.
- Salam, M. T., Lin, P. C., Avol, E. L., Gauderman, W. J., and Gilliland, F. D. (2007) Microsomal epoxide hydrolase, glutathione S-transferase P1, traffic and childhood asthma. *Thorax* **62**, 1050–1057.
- Scott, R. W., Bergman, B. L., Bajpai, A., Hersh, R. T., Rodriguez, H., Jones, B. N., Barreda, C., Watts, S., and Bakera, J. B. (1985) Protease nexin. Properties and a modified purification procedure. *J. Biol. Chem.* **260**, 7029–7034.
- Sethi, J. M., and Rochester, C. L. (2000) Smoking and chronic obstructive pulmonary disease. *Clin. Chest Med.* **21**, 67–86.
- Silverman, E. K., Palmer, L. J., Mosley, J. D., Barth, M., Senter, J. M., Brown, A., Drazen, J. M., Kwiatkowski, D. J., Chapman, H. A., Campbell, E. J., et al. (2002) Genomewide linkage analysis of quantitative spirometric phenotypes in severe early-onset chronic obstructive pulmonary disease. *Am. J. Hum. Genet.* **70**, 1229–1239.
- Snider, G. L. (1989) Chronic obstructive pulmonary disease: risk factors, pathophysiology and pathogenesis. *Annu. Rev. Med.* **40**, 411–429.
- Stenson, P. D., Ball, E., Howells, K., Phillips, A., Mort, M., and Cooper, D. N. (2008) Human Gene Mutation Database: towards a comprehensive central mutation database. *J. Med. Genet.* **45**, 124–126.
- Su, Z., Wen, F. Q., Feng, Y. L., Xiao, M., and Wu, X. L. (2005) Transforming growth factor-beta1 gene polymorphisms associated with chronic obstructive pulmonary disease in Chinese population. *Acta Pharmacol. Sin.* **26**, 714–720.
- van Diemen, C. C., Postma, D. S., Aulchenko, Y. S., Snijders, P. J., Oostra, B. A., van Duijn, C. M., and Boezen, H. M. (2010) Novel strategy to identify genetic risk factors for COPD severity: a genetic isolate. *Eur. Respir. J.* **35**, 768–775.
- Vaughan, P. J., and Cunningham, D. D. (1993) Regulation of protease nexin-1 synthesis and secretion in cultured brain cells by injury-related factors. *J. Biol. Chem.* **268**, 3720–3727.
- Vibhuti, A., Arif, E., Deepak, D., Singh, B., and Qadar, Pasha, M. A. (2007) Genetic polymorphisms of *GSTP1* and *mEPHX* correlate with oxidative stress markers and lung function in COPD. *Biochem. Biophys. Res. Commun.* **359**, 136–142.
- Wu, L., Chau, J., Young, R., Pokorny, V., Mills, G., Hopkins, R., McLean, L., and Black, P. (2004) Transforming growth factor-beta1 genotype and susceptibility to chronic obstructive pulmonary disease. *Thorax* **59**, 126–129.
- Xiao, D., Wang, C., Du, M. J., Pang, B. S., Zhang, H. Y., Xiao, B., Liu, J. Z., Weng, X., L, Su, L., and Christiani, D. C. (2003) Association between polymorphisms in the microsomal epoxide hydrolase (mEH) gene and chronic obstructive pulmonary disease. *Zhonghua Yi Xue Za Zhi*, **83** (20), 1782–1786.
- Zhao, T. M., and Lee, T. D. (1989) Gm and Km allotypes in 74 Chinese populations: a hypothesis of the origin of the Chinese nation. *Hum. Genet.* **83**, 101–110.
- Zhong, L., Fu, W.-P., Sun, C., Dai, L.-M., and Zhang, Y.-P. (2009) Absence of association between *SERPINE2* genetic polymorphisms and chronic obstructive pulmonary disease in Han Chinese: a case-control cohort study. *BMC Med. Genet.* **10**, 66–71.
- Zhu, G., Warren, L., Aponte, J., Gulsvik, A., Bakke, P., Anderson, W. H., Lomas, D. A., Silverman, E. K., and Pillai, S. G. (2007) The *SERPINE2* gene is associated with chronic obstructive pulmonary disease in two large populations. *Am. J. Respir. Crit. Care Med.* **176**, 167–173.
- Zhu, Y., Hoffman, A., Wu, X., Zhang, H., Zhang, Y., Leaderer, D., and Zheng, T. (2008) Correlating observed odds ratios from lung cancer case-control studies to SNP functional scores predicted by bioinformatic tool. *Mut. Res.* **639**, 80–88.