Research Article

Cancer Epidemiology, Biomarkers & Prevention

A Novel Pathway-Based Approach Improves Lung Cancer Risk Prediction Using Germline Genetic Variations

David C. Qian[1], Younghun Han[1], Jinyoung Byun[1], Hae Ri Shin[1], Rayjean J. Hung[2], John R. McLaughlin[3], Maria Teresa Landi[4], Daniela Seminara[4], and Christopher I. Amos[1]

Abstract

Background: Although genome-wide association studies (GWAS) have identified many genetic variants that are strongly associated with lung cancer, these variants have low penetrance and serve as poor predictors of lung cancer in individuals. We sought to increase the predictive value of germline variants by considering their cumulative effects in the context of biologic pathways.

Methods: For individuals in the Environment and Genetics in Lung Cancer Etiology study (1,815 cases/1,971 controls), we computed pathway-level susceptibility effects as the sum of relevant SNP variant alleles weighted by their log-additive effects from a separate lung cancer GWAS meta-analysis (7,766 cases/37,482 controls). Logistic regression models based on age, sex, smoking, genetic variants, and principal components of pathway effects and pathway–smoking interactions were trained and optimized in cross-validation and further tested on an independent dataset (556 cases/830 controls). We assessed prediction performance using area under the receiver operating characteristic curve (AUC).

Results: Compared with typical binomial prediction models that have epidemiologic predictors (AUC = 0.607) in addition to top GWAS variants (AUC = 0.617), our pathway-based smoking-interactive multinomial model significantly improved prediction performance in external validation (AUC = 0.656, $P < 0.0001$).

Conclusions: Our biologically informed approach demonstrated a larger increase in AUC over nongenetic counterpart models relative to previous approaches that incorporate variants.

Impact: This model is the first of its kind to evaluate lung cancer prediction using subtype-stratified genetic effects organized into pathways and interacted with smoking. We propose pathway–exposure interactions as a potentially powerful new contributor to risk inference. *Cancer Epidemiol Biomarkers Prev; 25(8); 1208–15. ©2016 AACR.*

Introduction

Lung cancer is the most common malignancy worldwide, with 1.3 million new cases per year. It is also the top cause of cancer-related deaths, responsible for 1.4 million deaths per year (1). Advanced-stage disease is commonly observed at diagnosis; 22% of newly diagnosed lung cancers have infiltrated regional lymph nodes and 57% have metastasized beyond the lung (2). The National Lung Screening Trial showed that among current smokers between 55 and 74 years of age with a smoking history of at least 30 pack-years, and analogous former smokers who quit for less than 15 years, lung cancer screening by low-dose helical CT confers a 20% reduction in mortality compared with screening by chest X-ray (3). This finding prompted the U.S. Preventive Services Task Force to recommend annual CT screening for such high-risk individuals (4). However, in practice, clinical pursuit of lung cancer usually occurs when patients already present with suggestive symptoms (5). Although the 5-year survival of overall lung cancer is 6% to 15%, that of stage I lung cancer is estimated to be 70%. Therefore, better identifying individuals with elevated lung cancer risk that motivates action may enable earlier diagnoses and improve outcomes (6).

The major lung cancer prediction models upon which most subsequent models (7) expand are the Bach model (8), the Spitz model (9), the Liverpool Lung Project (LLP) model (10), and the Tammemagi model (11). Although they differ slightly in population study design, analysis of smokers versus nonsmokers, and epidemiologic risk factors under consideration, all of the models agree that ascertaining airborne exposures and history of respiratory illnesses is useful for prediction. In contrast, extended models incorporating variants identified by lung cancer genome-wide association studies (GWAS) have so far demonstrated that germline markers contribute little to risk prediction due to their small effect sizes. For example, adding genome-wide significant SNPs rs8034191 and rs402710 increased area under the receiver operating characteristic curve (AUC) of the Bach model by 0.02 to 0.04 (separate analyses for current, former, and ever smokers). However, training and test sets were drawn from the same study for cross-validation (12). AUC increased by 0.03 with the addition of SNP rs663048 to the LLP model. However, predictions were tested on the same individuals from whom the model had been trained (13). The

[1]Department of Biomedical Data Science, Dartmouth Geisel School of Medicine, Lebanon, New Hampshire. [2]Lunenfeld-Tanenbaum Research Institute of Mount Sinai Hospital, Toronto, Canada. [3]Dalla Lana School of Public Health, University of Toronto, Toronto, Canada. [4]NCI, NIH, Bethesda, Maryland.

Note: Supplementary data for this article are available at Cancer Epidemiology, Biomarkers & Prevention Online (http://cebp.aacrjournals.org/).

Corresponding Author: Christopher I. Amos, Dartmouth Geisel School of Medicine, 1 Medical Center Drive, Williamson Translational Research Building, Lebanon, NH 03756. Phone: 603-650-1972; Fax: 603-653-6696; E-mail: Christopher.I.Amos@dartmouth.edu

doi: 10.1158/1055-9965.EPI-15-1318

Spitz model with a panel of inflammatory SNPs actually tested predictions on an independent study but observed a gain of only 0.01 in AUC (14). In a decision tree analysis with 6 SNPs, the 0.008 rise in AUC was not statistically significant ($P = 0.056$) in external validation (15). Another analysis constructed polygenic risk scores from genome-wide significant SNPs and improved AUC by 0.02 over the baseline nongenetic model in cross-validation (16).

We sought to enhance prediction by modeling genetic risk as the aggregate effects of disease-associated germline variants within biologic pathways. Subtype-specific discriminatory pathways for the two most common lung cancer subtypes, adenocarcinoma, and squamous cell carcinoma (SCC), as well as pathway–smoking interactions, were included to reflect potential histology-dependent and interactive influences as comprehensively as possible. We also derived variant effect sizes, trained our model, and tested our model using three independent datasets that were intentionally not pruned into a consensus set of SNPs to evaluate the generalizability of our approach.

Materials and Methods

Variant effect sizes from lung cancer GWAS meta-analyses

Fixed effects meta-analysis with inverse variance weighting was conducted using previously reported GWAS of 11,864,235 genotyped and imputed SNPs in 7,766 lung cancer cases (2,424 adenocarcinoma cases and 2,274 SCC cases) and 37,482 controls of European ancestry by the Transdisciplinary Research in Cancer of the Lung (TRICL) consortium (17–25), including stratified analyses for adenocarcinoma and SCC (Table 1). These data are available in the Database of Genotypes and Phenotypes (accession number phs000877.v1.p1).

Individual-level data for training and testing risk prediction models

IARC lung cancer study. Representative linkage disequilibrium structure (26) of the TRICL meta-analysis: The International Agency for Research on Cancer (IARC) study of lung cancer in central Europe recruited subjects from Czech Republic, Hungary, Poland, Romania, Russia, and Slovakia (22). After quality control (QC) defined below, 303,135 SNPs (Illumina HumanHap300) in 1,841 cases and 2,441 controls remain.

Table 1. Component studies of TRICL lung cancer GWAS meta-analyses

	Cases			
	All	AC[a]	SCC[b]	Controls
MDACC[c]	1,150	619	306	1,134
UKICR[d]	1,952	465	611	5,200
IARC[e]	2,533	517	911	3,791
Toronto[f]	331	90	90	499
Germany[g]	481	186	97	478
deCODE[h]	1,319	547	259	26,380
Totals	7,766	2,424	2,274	37,482

Abbreviation: AC, adenocarcinoma.
[a]Lung adenocarcinoma.
[b]Lung squamous cell carcinoma.
[c]MD Anderson Cancer Center study (18).
[d]United Kingdom Institute of Cancer Research study (19–21).
[e]International Agency for Research on Cancer study (22, 23).
[f]University of Toronto and Lunenfeld-Tanenbaum Research Institute study (22).
[g]Helmholtz-Gemeinschaft Deutscher Forschungszentren study (24).
[h]deCODE Genetics study (25).

Environment and Genetics in Lung Cancer Etiology study. Prediction training data: The Environment and Genetics in Lung Cancer Etiology (EAGLE) study investigated genetic and environmental determinants of lung cancer and smoking persistence in Italians (27). Age and smoking history were ascertained at the time of enrollment, which also coincided with age at lung cancer diagnosis for cases (incident cases). After QC, 501,658 SNPs (Illumina HumanHap550v3.0) in 1,815 cases and 1,971 controls remain.

The Prostate, Lung, Colorectal, and Ovarian Cancer screening trial. Prediction test data: Lung cancer cases were selected from both arms of the screening trial and were frequency matched by sex and age in 5-year intervals with controls from the lung and prostate components of the Prostate, Lung, Colorectal, and Ovarian Cancer (PLCO) screening trial, all European Americans (28). Age and smoking history were recorded at the time of intervention randomization for controls and lung cancer diagnosis for cases. After QC, 502,961 SNPs (Illumina HumanHap550v3.0) in 556 cases and 830 controls remain.

We excluded individuals with more than 10% missing genotypes or any missing age, sex, and smoking pack-years information, SNPs with minor allele frequency less than 5%, SNPs with genotyping rate less than 90%, and SNPs that failed the Hardy–Weinberg test at the 0.0001 significance level (Table 2). Genotype data were then imputed to the 1000 Genomes Project (phase III; ref. 29) with haplotype phasing by SHAPEIT (30) using IMPUTE2 v2.3.1 (31). We accepted the best-guess genotypes of imputed SNPs with information measure greater than 0.9. This stringent quality filter was applied because subsequent analyses require hard call genotypes, unlike in GWAS, where >0.3 to 0.4 is accepted to allow noninteger allelic dosages (32). If germline-based lung cancer risk assessment ever becomes a commercial reality with validated medical utility, it is doubtful that relevant genetic markers would be imputed with uncertainty from surrounding markers on a research chip. They would be directly genotyped on a disease-specific chip. We wanted to approximate such a scenario in our training and test datasets.

Statistical analyses

The study design overview is presented in Fig. 1. We trained logistic regression models on individuals in the EAGLE study to predict lung cancer status for individuals in the PLCO study. In

$$\ln \frac{P(Y_i = k)}{P(Y_i = \text{no cancer})} = \boldsymbol{\gamma} \mathrm{X} + \boldsymbol{\varepsilon},$$

Y_i is the disease status of individual i, having no lung cancer denotes the reference outcome, k an alternative outcome, $\mathbf{X}$ a matrix of predictor values, $\boldsymbol{\gamma}$ a vector of predictor effects, and $\boldsymbol{\varepsilon}$ a vector of error terms. Different choices for predictors are described below. In binomial logistic regression, k denotes having lung cancer. For a set of test instances, lung cancer prediction performance was evaluated by AUC applied to the range of risk prediction scores $P(Y_i = k)$. In multinomial logistic regression, k denotes having one of several possible lung cancer subtypes: "adenocarcinoma," "SCC," or "other lung cancer." "Other lung cancer" pools together lung cancer subtypes that are neither adenocarcinoma nor SCC, including mixed or unknown histology. Overall lung cancer prediction scores were computed as the linear combination of subtype prediction probabilities:

Table 2. Characteristics of training and test study populations

	EAGLE			PLCO		
	Cases, *n* (%)	**Controls, *n* (%)**	***P*[a]**	**Cases, *n* (%)**	**Controls, *n* (%)**	***P*[a]**
Total individuals	1,815	1,971		556	830	
Lung cancer subtype						
AC	753 (41.4)			267 (48.0)		
SCC	466 (25.7)			122 (22.0)		
Other	596 (32.8)			167 (30.0)		
Sex			0.118			0.169
Male	1,429 (78.7)	1,509 (76.6)		341 (61.3)	477 (57.5)	
Female	386 (21.3)	462 (23.4)		215 (38.7)	353 (42.5)	
Age			0.023			0.238
≤59	395 (21.8)	502 (25.5)		106 (19.1)	177 (21.3)	
60–64	316 (17.4)	349 (17.7)		167 (30.0)	262 (31.6)	
65–69	406 (22.4)	451 (22.9)		178 (32.0)	267 (32.2)	
70–74	399 (22.0)	400 (20.3)		105 (18.9)	124 (14.9)	
≥75	299 (16.5)	269 (13.6)		0 (0.0)	0 (0.0)	
Cumulative pack-years smoked			$<1 \times 10^{-10}$			$<1 \times 10^{-10}$
0	138 (7.6)	633 (32.1)		51 (9.2)	75 (9.0)	
1–15	123 (6.8)	468 (23.7)		69 (12.4)	231 (27.8)	
16–30	278 (15.3)	356 (18.1)		74 (13.3)	109 (13.1)	
31–40	306 (16.9)	194 (9.8)		106 (19.1)	168 (20.2)	
41–50	320 (17.6)	149 (7.6)		51 (9.2)	64 (7.7)	
51–60	251 (13.8)	81 (4.1)		54 (9.7)	54 (6.5)	
61–70	120 (6.6)	29 (1.5)		46 (8.3)	42 (5.1)	
71–80	85 (4.7)	30 (1.5)		28 (5.0)	23 (2.8)	
≥81	194 (10.7)	31 (1.6)		77 (13.8)	64 (7.7)	
Genetic data						
Genotyped SNPs		501,658			502,961	
Imputed SNPs		6,652,756			6,535,115	

Abbreviation: AC, adenocarcinoma.

[a]Calculated using the χ^2 test to compare sex and the *t* test to compare age and smoking pack-years.

$a \cdot P(\text{AC}) + b \cdot P(\text{SCC}) + c \cdot P(\text{other})$. Over a grid search on [0.01, 4] in 0.01 increments, optimal a^*, b^*, c^* were identified through 400^3 rounds of 5-fold internal cross-validation within the EAGLE study. Then following model training on the entire EAGLE study, lung cancer prediction performance was evaluated by AUC applied to the range of risk prediction scores $S = a^* \cdot P(\text{AC}) + b^* \cdot P(\text{SCC}) + c^* \cdot P(\text{other})$ for individuals in the PLCO study. Mean, confidence interval, and comparison tests by DeLong method (33) of AUC were computed using the R package "pROC."

Epidemiologic factors, age, sex, and smoking pack-years. Equation 1 was fit to lung cancer status, age, sex, and smoking pack-years of individuals in the EAGLE study and used to infer lung cancer status of individuals in the PLCO study. In this model and subsequent models, the increases in log odds of having outcome k relative to not having lung cancer with every additional year of age, being female, and every additional pack-year of smoking are $\beta_{1,k}$, $\beta_{2,k}$, and $\beta_{3,k}$, respectively.

$$\ln\frac{P(Y_i = k)}{P(Y_i = \text{no cancer})} = \alpha_k + \beta_{1,k} \cdot age_i + \beta_{2,k} \cdot sex_i + \beta_{3,k} \cdot packyears_i \quad (1)$$

Epidemiologic factors and top GWAS SNPs. We applied the software Genome-Wide Complex Trait Analysis (GCTA; ref. 34) to summary statistics of the TRICL meta-analysis to determine independently associated SNPs through stepwise-selection conditional analysis (26). In the absence of individual-level genotype data for the meta-analysis, GCTA estimated linkage disequilibrium structure from the IARC study. Of the 1,343 SNPs deemed to be independently associated (GWAS $P < 0.05$ and conditional association $P < 0.001$), 301 SNPs appear in both EAGLE and PLCO (Supplementary Table S1). Using the EAGLE study, equation 2 was fit to epidemiologic factors plus the top n independently associated SNP(s) ranked by ascending GWAS P value ($1 \leq n \leq 301$) and was tested on the PLCO study. We reported the highest AUC for each set of 301 binomial and multinomial models. The increase in relative log odds with every additional copy of the variant allele for SNP j is $\theta_{j,k}$.

$$\ln\frac{P(Y_i = k)}{P(Y_i = \text{no cancer})} = \alpha_k + \beta_{1,k} \cdot age_i + \beta_{2,k} \cdot sex_i + \beta_{3,k} \cdot packyears_i + \sum_{j=1}^{n} \theta_{j,k} \cdot SNP_{j,i} \quad (2)$$

Epidemiologic factors, top GWAS SNPs, and SNP–smoking interactions. Corresponding smoking interactions of the top n independently associated SNPs were added to the previous model. The increases in relative log odds with every additional pack-year of smoking and copy of the variant allele for SNP j are $\beta_{3,k} + \sum_{j=1}^{n} \xi_{j,k} \cdot SNP_{j,i}$ and $\theta_{j,k} + \xi_{j,k} \cdot packyears_i$, respectively.

$$\ln\frac{P(Y_i = k)}{P(Y_i = \text{no cancer})} = \alpha_k + \beta_{1,k} \cdot age_i + \beta_{2,k} \cdot sex_i + \beta_{3,k} \cdot packyears_i + \sum_{j=1}^{n} \theta_{j,k} \cdot SNP_{j,i} + \sum_{j=1}^{n} \xi_{j,k} \cdot SNP_{j,i} \cdot packyears_i \quad (3)$$

Epidemiologic factors and biologic pathways. For nominally significant SNPs in the TRICL meta-analysis (association $P < 0.05$) that may affect a gene through either residing within the gene or

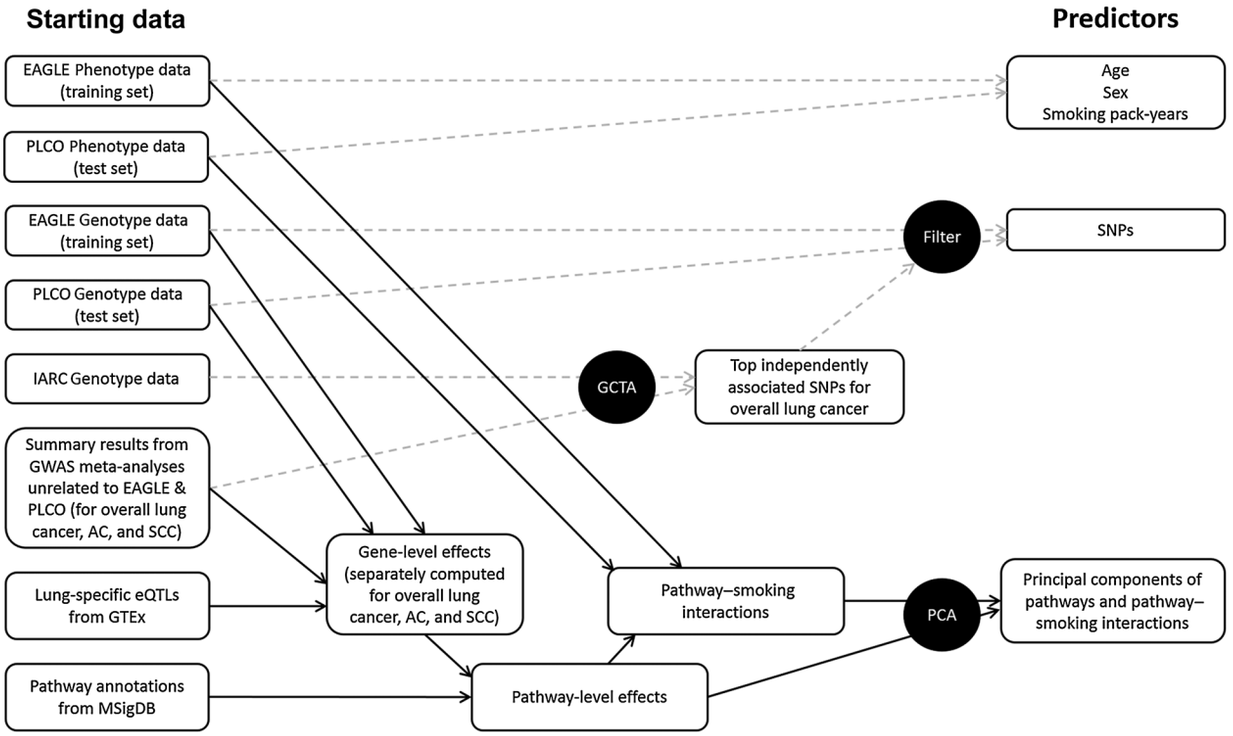

Figure 1.

Schematic overview of study design. Procedures for deriving lung cancer risk predictors from the phenotypes and genotypes of independent datasets are outlined. Dashed arrows refer to methods that have been previously reported or that are similar to those previously reported. Bold arrows refer to new approaches reported herein. AC, adenocarcinoma; GTEx, Genotype-Tissue Expression project; PCA, principal component analysis; MSigDB, Molecular Signatures Database.

influencing the gene's expression in lung [expression quantitative loci (eQTL) FDR < 0.05 in the Genotype-Tissue Expression project; ref. 35], we determined the number of variant alleles possessed by individuals in the EAGLE and PLCO studies. Although some SNPs are available in one study but not the other, the following construction of pathway predictors does not depend on identical genotype data between the two studies. Gene-level susceptibility effects were first computed for every individual in both studies as variant allele count (0, 1, or 2) multiplied by additive effect size (log OR) of each gene's lung cancer-associated mapped SNP from the meta-analysis. We then derived pathway-level susceptibility effects for all individuals with respect to every curated pathway in the Molecular Signatures Database (36) as the sum of effects for genes relevant to that pathway. Using a randomly sampled 80% of the EAGLE study, we identified pathways that exhibit significant differences (t test $P < 0.05$) between lung cancer cases and controls. Principal components (PC) of pathway z-scores were separately derived for the internal training and testing data (remaining 20% of the EAGLE study) using the R package "FactoMineR." A binomial logistic regression model was then fit to epidemiologic factors plus the top p PCs of the 80% EAGLE training set and tested on the 20% EAGLE test set. This process was repeated 30×100 times for positive integers $p \leq 30$ and across one hundred 80/20 resamplings of EAGLE to optimize AUC. Using the optimal p^*, we trained equation 4 on the entire EAGLE study to infer lung cancer status in the PLCO study. The increase in relative log odds with every unit increment of PC x is $\phi_{x,k}$.

$$\ln \frac{P(Y_i = k)}{P(Y_i = \text{no cancer})} = \alpha_k + \beta_{1,k} \cdot age_i + \beta_{2,k} \cdot sex_i + \beta_{3,k} \cdot packyears_i + \sum_{x=1}^{p^*} \phi_{x,k} \cdot PC_{x,i} \qquad (4)$$

Epidemiologic factors, biologic pathways, and pathway–smoking interactions. We repeated the previous analysis, except PCs were derived for both pathways and pathway–smoking interactions (Supplementary Table S2). Pathway–smoking interactions were constructed as the product of pathway effects and corresponding cumulative smoking pack-years.

Epidemiologic factors, subtype-specific biologic pathways, and subtype-specific pathway–smoking interactions. We repeated the previous analysis, except with pathways and pathway–smoking interactions from stratified GWAS meta-analyses for adenocarcinoma and SCC (Supplementary Tables S3 and S4), in addition to those for overall lung cancer. With respect to the top p overall lung cancer PCs, top q adenocarcinoma PCs, and top r SCC PCs, $30^3 \times 100$ rounds of internal cross-validation for positive integers $p, q, r \leq 30$ across one hundred 80/20 resamplings of EAGLE were performed to optimize AUC. Solutions p^*, q^*, r^* were then used to train equation 5 on the entire EAGLE study for prediction in the PLCO study. The increases in relative log odds with every unit

Table 3. Performance of lung cancer prediction models

Logistic regression model	AUC (95% CI, *P*)[a]	NRI (95% CI, *P*)[b]	Cohen κ (95% CI)
1) Epidemiologic: age, sex, PY			
Binomial	0.607 (0.577–0.637)		0.172 (0.120–0.224)
Multinomial	0.608 (0.577–0.638)		0.081 (0.037–0.125)
2) Epidemiologic + top independently associated SNPs[c]			
Binomial	0.617 (0.587–0.647, ref.)	(Ref.)	0.194 (0.141–0.246)
Multinomial	0.617 (0.587–0.647, 0.618)	3.6% (0.9–6.4, 0.0104)	0.095 (0.047–0.144)
3) Epidemiologic + top independently associated SNPs[c] and their interactions with PY			
Binomial	0.619 (0.589–0.650, 0.598)	2.9% (0.7–5.1, 0.0097)	0.195 (0.143–0.247)
Multinomial	0.620 (0.590–0.650, 0.586)	2.2% (0.1–4.3, 0.0438)	0.101 (0.052–0.150)
4) Epidemiologic + PCs of top overall pathways			
Binomial	0.621 (0.591–0.651, 0.568)	5.3% (1.8–8.8, 0.0027)	0.200 (0.146–0.253)
Multinomial	0.621 (0.591–0.651, 0.582)	5.0% (1.7–8.4, 0.0033)	0.105 (0.058–0.152)
5) Epidemiologic + PCs of top overall pathways and their interactions with PY			
Binomial	0.630 (0.600–0.660, 3.28×10^{-2})	4.9% (2.5–7.4, 0.0002)	0.207 (0.154–0.261)
Multinomial	0.631 (0.600–0.661, 2.94×10^{-2})	5.0% (2.4–7.6, 0.0002)	0.125 (0.077–0.173)
6) Epidemiologic + PCs of top subtype-specific pathways and their interactions with PY			
Binomial	0.651 (0.621–0.681, 8.78×10^{-4})	8.9% (5.7–12.1, <0.0001)	0.226 (0.171–0.281)
Multinomial	0.656 (0.626–0.685, 6.11×10^{-5})	11.7% (7.3–16.0, <0.0001)	0.152 (0.108–0.195)

Abbreviations: CI, confidence interval; PY, pack-years.

[a]Changes in AUC were computed relative to the #2 binomial model as reference.

[b]Net reclassification improvement values (%) were computed relative to the #2 binomial model as reference. Classification was determined on the basis of choosing a risk prediction score cutoff for each model that corresponds to the ROC curve point with minimum Euclidean distance to coordinate (0, 1).

[c]A total of 301 independently associated SNPs were identified by stepwise selection conditional analysis of the TRICL overall lung cancer GWAS meta-analysis. Logistic regression models that include the top *n* SNPs were tested, for $1 \le n \le 301$. Optimal *n* yielding the highest AUC were 30 and 39 for models 2 and 3, respectively.

increment of PCs *x*, *y*, and *z* for overall lung cancer, adenocarcinoma, and SCC are $\phi_{x,k}$, $\psi_{y,k}$, and $\omega_{z,k}$, respectively.

$$\ln\frac{P(Y_i=k)}{P(Y_i=\text{no cancer})} = \alpha_k + \beta_{1,k}\cdot age_i + \beta_{2,k}\cdot sex_i + \beta_{3,k}\cdot packyears_i + \sum_{x=1}^{p^*}\phi_{x,k}\cdot PC_{x,i}^{\text{overall}} + \sum_{y=1}^{q^*}\psi_{y,k}\cdot PC_{y,i}^{\text{AC}} + \sum_{z=1}^{r^*}\omega_{z,k}\cdot PC_{z,i}^{\text{SCC}} \quad (5)$$

Classification accuracy. We computed net reclassification improvement (NRI; ref. 37) for the pathway-based models compared with the SNP-based binomial model as reference, using the R package "PredictABEL" (Table 3). The decision score cutoff for each model was chosen as the ROC curve point with minimum Euclidean distance to coordinate (0, 1). This cutoff was also used to compute Cohen κ (38) between predicted and observed outcomes (no cancer or lung cancer); it is a measure of concordance beyond what would be expected by chance alone, ranging from 0 (concordance due to chance) to 1 (perfect concordance). Likewise for the multinomial models, a predicted disease subtype (no cancer, adenocarcinoma, SCC, or other lung cancer) for each individual was chosen as the one with highest probability and Cohen κ was computed.

Results

To assess differences in the distribution of demographic variables between cases and controls in the EAGLE and PLCO studies, we used the χ^2 test to compare categorical variables and *t* test to compare continuous variables (Table 2). Cases were much more likely to have longer smoking histories. Lung cancer exhibits modest association with age in the EAGLE study, but not in the PLCO study.

Logistic regression models were trained on the EAGLE study and tested on the PLCO study. A baseline binomial model with only age, sex, and smoking pack-years as predictors attained an AUC of 0.607 (#1 in Table 3). Consistent with previous findings (7), adding top SNPs implicated in overall lung cancer GWAS hardly increased AUC, by around 0.01 (#2 in Table 3). Diminishing returns came from further adding smoking interactions with these SNPs (#3 in Table 3). To our knowledge, existing studies of genetic interactions with smoking in lung cancer tend to be discovery in nature and have not evaluated risk prediction (14, 39–41). Hence, comparable AUC gains are unavailable. While incorporating top PCs of discriminatory pathway-level effects from overall lung cancer GWAS negligibly boosted AUC as well (#4 in Table 3), significant improvements were achieved by including pathway–smoking interactions along with pathways derived from adenocarcinoma and SCC subtype-specific GWAS (#s 5 and 6 in Table 3).

AUC is usually favored in the early phase of prediction modeling because setting decision cutoffs is not required. However, a common critique by clinical researchers is that a specific cutoff must be selected to assess patient impact (42). We chose the lung cancer prediction score cutoff for NRI that equally values sensitivity and specificity. Compared with the SNP-based binomial model, our best model correctly net reclassified 11.7% of individuals in the PLCO study. Sensitivity improved from 56% to 66%, and specificity improved from 61% to 63%. A recent external validation study of several popular nongenetic lung cancer prediction models, also treating true positives and false positives as equal tradeoffs, yielded sensitivities all less than 66% despite having incorporated many more epidemiologic risk factors, such as asbestos exposure, COPD, hay fever, and family history of lung cancer (43). Our model's specificity was not as high though. These findings suggest that genetic considerations may be more important for identifying cases, whereas epidemiologic considerations may be more important for identifying controls. Genetics-informed lung cancer prediction models have reported AUCs, but not

sensitivities and specificities at discrete ROC curve cutoffs against which to further compare (7, 12–16).

Multinomial logistic regressions assigned lung cancer subtypes to individuals in the PLCO study. Although Cohen κ of the most comprehensive model is nearly double that of the nongenetic model, classification accuracies across multinomial models are low in absolute terms. As expected, all of the binomial models exhibit higher κ values, mainly due to the difference in difficulty of predicting 2 outcomes versus 4 outcomes. Overall lung cancer scores generated by merging lung cancer subtype probabilities from the multinomial models also achieved AUCs similar to those of counterpart binomial models. However, the value of the multinomial approach becomes apparent when examining ordered lung cancer prediction scores. With individuals in the bottom decile of scores from the best multinomial model (#6 in Table 3) as reference, those in the upper deciles have ORs of developing lung cancer that follow a more evenly graded progression with a wider spread, compared with predicted risk deciles from the partner binomial model. In particular, the difference in stratification at the extremes is highly significant (Fig. 2). On the basis of the multinomial model, odds of developing lung cancer for PLCO subjects in the top predicted risk decile is 4.68 times the odds in the bottom decile. The binomial model was not as proficient in identifying the highest and lowest risk individuals, producing a corresponding OR of 3.42.

Beyond these deciles, constructing clinically meaningful risk categories is outside the scope of this study. Nevertheless, we still identified individuals at high risk for lung cancer in a way that could at least be compared with the U.S. Preventive Services Task Force criteria: adults aged 55 to 80 years with at least 30 pack-years of smoking history, including former smokers who quit for less than 15 years (4). Of the 1,386 individuals in the PLCO test set, 777 meet this criteria and 362 (47%) actually developed lung cancer. On the other hand, the 777 highest scoring individuals in PLCO from the #6 multinomial model consist of 435 (56%) lung cancer cases, yielding a positive predictive value increase of 9%.

Discussion

We put forward a method of improving lung cancer risk prediction that is innovative in several ways. As lung cancer is a heterogeneous disease, we used SNP effect sizes from overall lung cancer and subtype-stratified GWAS. Although selection of SNP predictors has traditionally involved filtering SNPs by GWAS *P* value or suspected functions (12–16), we aggregated relevant SNP effects into pathway-level effects. Many studies have already warned about the shortcomings of conducting mainly biostatistical polygenic analyses of GWAS for complex disease prediction (44), even if a larger than expected portion of "missing heritability" has been extrapolated to be hidden rather than actually missing (34). Our biologic integration also included constructing pathway–smoking interactions, the first use of pathway–exposure interactions in germline genetic prediction.

SNP–gene correspondences were established on the basis of intragenic or lung eQTL status, a refinement of the usual proximity approach. As an exercise, we recomputed pathway effects by mapping SNPs to genes within 10 kilobases (kb) and discovered that the differences between lung cancer cases and controls became less pronounced (data not shown). This suggests previous pathway-based analyses of GWAS that mapped SNPs to genes based solely on chromosomal position may have introduced noise. Not all genetic variations within 10 kb of a gene influence its expression, and some variants farther than 10 kb away can also influence expression. Pathway values derived using the position method did not distinguish cases from controls as strikingly because the sums of gene-level effects may have included genes that were improperly linked to disease-associated SNPs from GWAS.

Furthermore, simultaneous consideration of lung cancer subtypes through multinomial logistic regression identified individuals with the highest and lowest lung cancer risk better than binomial regression. This finding is in line with a study that demonstrated jointly predicting risk for schizophrenia, bipolar disorder, and depression using SNP data is more effective than making separate predictions (45). It is believed that genetic

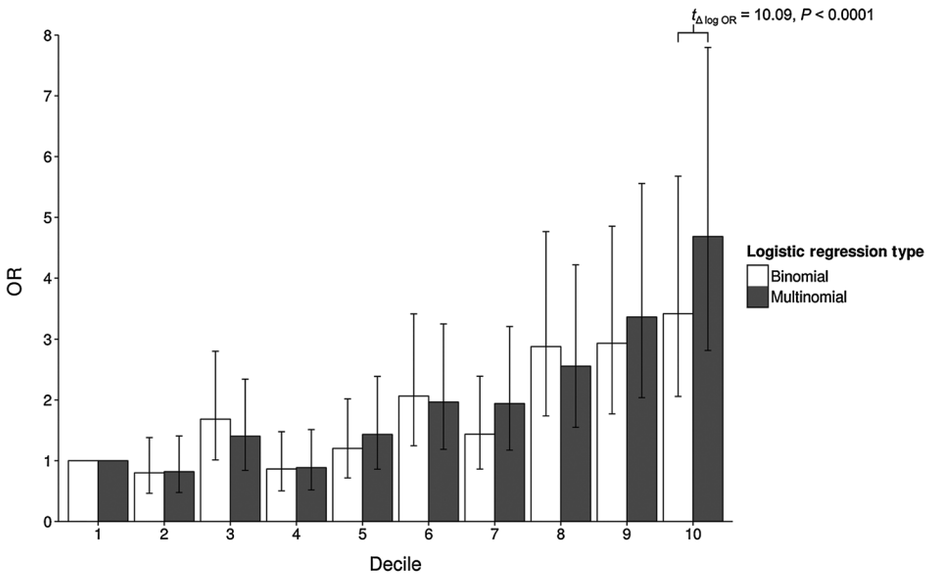

Figure 2.
ORs by risk prediction score deciles. Individuals in the PLCO study were grouped into deciles based on their risk prediction scores from the #6 binomial and multinomial models in Table 3. The odds of having lung cancer for each decile were compared with the odds for the lowest decile. Bars denote the 95% confidence interval for these ORs.

correlation among several diseases implies a variant affecting risk for one disease will tend to be informative of risk for correlated diseases as well. However, κ values from joint disease prediction are not impressive. Similar to GWAS of most complex diseases, many associated genetic risk loci exist for lung cancer, and their effect sizes have been measured to be quite small and with noise (44). Therefore, prediction accuracy at the single disease (subtype) level is similarly poor. Nevertheless, merging subtype risks following multinomial analysis likely enhanced the signal-to-noise ratio of genetic effects (46) and has contributed to the more clinically important endeavor of better assessing overall lung cancer risk. After all, motivation for lung cancer screening does not depend on probable cancer subtype.

Incorporating subtype-specific pathways and pathway–smoking interactions increased AUC of overall lung cancer prediction by 0.05 and 0.04 over the baseline nongenetic model and the baseline model plus top independently associated SNPs, respectively, in external validation. Primarily limited by available data, our best model achieved an AUC of 0.656. Although this AUC is far from being qualified to sway individual clinical decisions, we emphasize the significance of our gain, rather than the absolute value of AUC. The current, most thorough nongenetic models that assess many other predictors in addition to age, sex, and smoking pack-years, such as history of respiratory illnesses, occupational exposures, and level of education, have achieved AUC of 0.797 in external validation (47). This performance can be expected to advance another 0.03 to 0.05 upon inclusion of pathway-based and interactive predictors because pathway effects are largely uncorrelated with epidemiologic risk factors (Supplementary Tables S2–S4). The projected attenuation is attributed to potential overlapping genetic effects on disease and other epidemiologic factors not ascertained here. Such a gain would bring us ever closer the goal of accurately identifying high-risk individuals for early lung cancer screening.

A drawback of this study is difficulty in interpreting the biologic impacts of genetic factors on risk prediction. From the training data, we derived pathways and pathway–smoking interactions that discern lung cancer cases from healthy controls (Supplementary Tables S2–S4). Of note, top pathways for adenocarcinoma and SCC are mostly different, and some pathway–smoking interactions are more significant than their corresponding non-interacted pathways. Prominent examples include mechanisms of Bcl11b (a zinc finger protein) in adenocarcinoma and regulation of Smad2/3/4 (downstream transducers of signaling by transforming growth factor β) in SCC. However, the precise risk-influencing effects of pathways and their constituent genes are masked by PC regression. Coefficient estimates cannot be ascribed to any single factor from the original data. This masking through orthogonalization somewhat absolves the concern that evaluating so many interactive terms would inevitably lead to false positive biologic inferences. The rankings presented in Supplementary Tables S2–S4 may still merit new experimental pursuits; actions of Bcl11b and Smad4 have been implicated in tumorigenesis of colon adenocarcinoma (48) and head and neck SCC (49), respectively. In addition, transforming to PCs has allowed removal of multicollinearity among many related pathways, an inherent feature of modern pathway databases, without discarding pathways altogether as the subtle distinctions may be important.

Another weakness of this study, along with all existing lung cancer prediction efforts, is poor ability to model risk for never-smokers. They lack the top risk factor for lung cancer and their pathway–smoking interactions are all zero. Adding pathways did not significantly augment AUC among PLCO never-smokers compared with the baseline nongenetic model (data not shown). With additional information and the adaptability of our pathway–exposure interactions, however, this inadequacy has the potential to change in future studies. Other exposures besides smoking, such as second-hand smoke, pollution, asbestos, various dusts, and diet, can increase risk for lung cancer as well (50). Modeling interactions between a variety of exposure risk factors and genetic pathway effects, especially pathways for adenocarcinoma as it is the major tumor subtype among never-smokers (51), may reveal new insights.

Disclosure of Potential Conflicts of Interest

No potential conflicts of interest were disclosed.

Authors' Contributions

Conception and design: D.C. Qian, D. Seminara, C.I. Amos
Development of methodology: D.C. Qian, J.R. McLaughlin, C.I. Amos
Acquisition of data (provided animals, acquired and managed patients, provided facilities, etc.): R.J. Hung, J.R. McLaughlin, M.T. Landi, D. Seminara, C.I. Amos
Analysis and interpretation of data (e.g., statistical analysis, biostatistics, computational analysis): D.C. Qian, Y. Han, J. Byun, H.R. Shin, C.I. Amos
Writing, review, and/or revision of the manuscript: D.C. Qian, R.J. Hung, J.R. McLaughlin, M.T. Landi, D. Seminara, C.I. Amos
Administrative, technical, or material support (i.e., reporting or organizing data, constructing databases): H.R. Shin, J.R. McLaughlin, C.I. Amos
Study supervision: J.R. McLaughlin, D. Seminara

Acknowledgments

The authors thank all members of the Transdisciplinary Research in Cancer of the Lung consortium.

Grant Support

This research was supported by the NIH (P30CA023108, U19CA148127, R01CA149462, and P20GM103534; to C.I. Amos) and the National Science Foundation Graduate-K12 Fellowship in collaboration with Kimball Union Academy (DGE-0947790; to D.C. Qian).

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked *advertisement* in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

Received December 30, 2015; revised May 12, 2016; accepted May 13, 2016; published OnlineFirst May 24, 2016.

References

1. Fitzmaurice C, Dicker D, Pain A, Hamavid H, Moradi-Lakeh M, MacIntyre MF, et al. The global burden of cancer 2013. JAMA Oncol 2015;1:505–27.
2. National Cancer Institute. SEER Stat Fact Sheets: Lung and Bronchus Cancer. Bethesda, MD: National Cancer Institute; 2015.
3. The National Lung Screening Trial Research Team, Aberle DR, Adams AM, Berg CD, Black WC, Clapp JD, et al. Reduced lung-cancer mortality with low-dose computed tomographic screening. N Engl J Med 2011;365:395–409.
4. Humphrey L, Deffebach M, Pappas M, Baumann C, Artis K, Priest Mitchell J, et al. Screening for lung cancer: systematic review to update the U.S. Preventive Services Task Force recommendation. Evidence synthesis no. 105. Rockville, MD: Agency for Healthcare Research and Quality; 2013.
5. Chen X, Gorlov IP, Ying J, Merriman KW, Kimmel M, Lu C, et al. Initial medical attention on patients with early-stage non-small cell lung cancer. PLoS One 2012;7:e32644.

6. Field JK, Chen Y, Marcus MW, McRonald FE, Raji OY, Duffy SW. The contribution of risk prediction models to early detection of lung cancer. J Surg Oncol 2013;108:304–11.
7. Wang X, Oldani MJ, Zhao X, Huang X, Qian D. A review of cancer risk prediction models with genetic variants. Cancer Inform 2014;13:19–28.
8. Bach PB, Kattan MW, Thornquist MD, Kris MG, Tate RC, Barnett MJ, et al. Variations in lung cancer risk among smokers. J Natl Cancer Inst 2003;95:470–8.
9. Spitz MR, Hong WK, Amos CI, Wu X, Schabath MB, Dong Q, et al. A risk model for prediction of lung cancer. J Natl Cancer Inst 2007;99:715–26.
10. Cassidy A, Myles JP, van Tongeren M, Page RD, Liloglou T, Duffy SW, et al. The LLP risk model: an individual risk prediction model for lung cancer. Br J Cancer 2008;98:270–6.
11. Tammemagi CM, Pinsky PF, Caporaso NE, Kvale PA, Hocking WG, Church TR, et al. Lung cancer risk prediction: prostate, lung, colorectal and ovarian cancer screening trial models and validation. J Natl Cancer Inst 2011;103: 1058–68.
12. Hoggart C, Brennan P, Tjonneland A, Vogel U, Overvad K, Ostergaard JN, et al. A risk model for lung cancer incidence. Cancer Prev Res 2012;5:834–46.
13. Raji OY, Agbaje OF, Duffy SW, Cassidy A, Field JK. Incorporation of a genetic factor into an epidemiologic model for prediction of individual risk of lung cancer: the Liverpool Lung Project. Cancer Prev Res 2010;3:664–9.
14. Spitz MR, Amos CI, Land S, Wu X, Dong Q, Wenzlaff AS, et al. Role of selected genetic variants in lung cancer risk in African Americans. J Thorac Oncol 2013;8:391–7.
15. Weissfeld JL, Lin Y, Lin HM, Kurland BF, Wilson DO, Fuhrman CR, et al. Lung cancer risk prediction using common SNPs located in GWAS-identified susceptibility regions. J Thorac Oncol 2015:10;1538–45.
16. Li H, Yang L, Zhao X, Wang J, Qian J, Chen H, et al. Prediction of lung cancer risk in a Chinese population using a multifactorial genetic model. BMC Med Genet 2012;13:118.
17. Timofeeva MN, Hung RJ, Rafnar T, Christiani DC, Field JK, Bickeboller H, et al. Influence of common genetic variation on lung cancer risk: meta-analysis of 14 900 cases and 29 485 controls. Hum Mol Genet 2012;21:4980–95.
18. Amos CI, Wu X, Broderick P, Gorlov IP, Gu J, Eisen T, et al. Genome-wide association scan of tag SNPs identifies a susceptibility locus for lung cancer at 15q25.1. Nat Genet 2008;40:616–22.
19. Wang Y, Broderick P, Webb E, Wu X, Vijayakrishnan J, Matakidou A, et al. Common 5p15.33 and 6p21.33 variants influence lung cancer risk. Nat Genet 2008;40:1407–9.
20. Eisen T, Matakidou A, Houlston R, GELCAPS Consortium. Identification of low penetrance alleles for lung cancer: the GEnetic Lung CAncer Predisposition Study (GELCAPS). BMC Cancer 2008;8:244.
21. Power C, Elliott J. Cohort profile: 1958 British birth cohort (National Child Development Study). Int J Epidemiol 2006;35:34–41.
22. Hung RJ, McKay JD, Gaborieau V, Boffetta P, Hashibe M, Zaridze D, et al. A susceptibility locus for lung cancer maps to nicotinic acetylcholine receptor subunit genes on 15q25. Nature 2008;452:633–7.
23. Scelo G, Constantinescu V, Csiki I, Zaridze D, Szeszenia-Dabrowska N, Rudnai P, et al. Occupational exposure to vinyl chloride, acrylonitrile and styrene and lung cancer risk (Europe). Cancer Causes Control 2004;15: 445–52.
24. Sauter W, Rosenberger A, Beckmann L, Kropp S, Mittelstrass K, Timofeeva M, et al. Matrix metalloproteinase 1 (MMP1) is associated with early-onset lung cancer. Cancer Epidemiol Biomarkers Prev 2008;17:1127–35.
25. Thorgeirsson TE, Geller F, Sulem P, Rafnar T, Wiste A, Magnusson KP, et al. A variant associated with nicotine dependence, lung cancer and peripheral arterial disease. Nature 2008;452:638–42.
26. Yang J, Ferreira T, Morris AP, Medland SE Genetic Investigation of ANthropometric Traits (GIANT) Consortium, DIAbetes Genetics Replication And Meta-analysis (DIAGRAM) Consortium, et al. Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. Nat Genet 2012;44:369–75.
27. Landi MT, Consonni D, Rotunno M, Bergen AW, Goldstein AM, Lubin JH, et al. Environment And Genetics in Lung cancer Etiology (EAGLE) study: an integrative population-based case-control study of lung cancer. BMC Public Health 2008;8:203.
28. Hayes RB, Sigurdson A, Moore L, Peters U, Huang WY, Pinsky P, et al. Methods for etiologic and early marker investigations in the PLCO trial. Mutat Res 2005;592:147–54.
29. Delaneau O, Marchini JThe 1000 Genomes Project Consortium. Integrating sequence and array data to create an improved 1000 Genomes Project haplotype reference panel. Nat Commun 2014;5:3934.
30. Delaneau O, Marchini J, Zagury JF. A linear complexity phasing method for thousands of genomes. Nat Methods 2012;9:179–81.
31. Howie BN, Donnelly P, Marchini J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS Genet 2009;5:e1000529.
32. Wang Y, McKay JD, Rafnar T, Wang Z, Timofeeva MN, Broderick P, et al. Rare variants of large effect in BRCA2 and CHEK2 affect risk of lung cancer. Nat Genet 2014;46:736–41.
33. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 1988;44:837–45.
34. Yang J, Lee SH, Goddard ME, Visscher PM. GCTA: a tool for genome-wide complex trait analysis. Am J Hum Genet 2011;88:76–82.
35. The GTEx Consortium. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. Science 2015;348:648–60.
36. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci 2005;102:15545–50.
37. Pencina MJ, D'Agostino RB, Steyerberg EW. Extensions of net reclassification improvement calculations to measure usefulness of new biomarkers. Stat Med 2011;30:11–21.
38. Cohen J. A coefficient of agreement for nominal scales. Educ Psychol Meas 1960;20:37–46.
39. VanderWeele TJ, Asomaning K, Tchetgen Tchetgen EJ, Han Y, Spitz MR, Shete S, et al. Genetic variants on 15q25.1, smoking, and lung cancer: an assessment of mediation and interaction. Am J Epidemiol 2012;175: 1013–20.
40. Rudd MF, Webb EL, Matakidou A, Sellick GS, Williams RD, Bridle H, et al. Variants in the GH-IGF axis confer susceptibility to lung cancer. Genome Res 2006;16:693–701.
41. Xun X, Wang H, Yang H, Wang H, Wang B, Kang L, et al. CLPTM1L genetic polymorphisms and interaction with smoking and alcohol drinking in lung cancer risk: a case-control study in the Han population from northwest China. Medicine 2014;93:e289.
42. Leening MJ, Vedder MM, Witteman JC, Pencina MJ, Steyerberg EW. Net reclassification improvement: computation, interpretation, and controversies: a literature review and clinician's guide. Ann Intern Med 2014;160: 122–31.
43. Li K, Husing A, Sookthai D, Bergmann M, Boeing H, Becker N, et al. Selecting high-risk individuals for lung cancer screening: a prospective evaluation of existing risk models and eligibility criteria in the German EPIC cohort. Cancer Prev Res 2015;8:777–85.
44. Wray NR, Yang J, Hayes BJ, Price AL, Goddard ME, Visscher PM. Pitfalls of predicting complex traits from SNPs. Nat Rev Genet 2013; 14:507–15.
45. Maier R, Moser G, Chen GB, Ripke S, Cross-Disorder Working Group of the Psychiatric Genomics Consortium, Coryell W, et al. Joint analysis of psychiatric disorders increases accuracy of risk prediction for schizophrenia, bipolar disorder, and major depressive disorder. Am J Hum Genet 2015;96:283–94.
46. Guo G, Zhao F, Wang Y, Zhang Y, Du L, Su G. Comparison of single-trait and multiple-trait genomic prediction models. BMC Genet 2014;15:30.
47. Tammemagi MC, Katki HA, Hocking WG, Church TR, Caporaso N, Kvale PA, et al. Selection criteria for lung-cancer screening. N Engl J Med 2013;368:728–36.
48. Sakamaki A, Katsuragi Y, Otsuka K, Tomita M, Obata M, Iwasaki T, et al. Bcl11b SWI/SNF-complex subunit modulates intestinal adenoma and regeneration after gamma-irradiation through Wnt/beta-catenin pathway. Carcinogenesis 2015;36:622–31.
49. Korc M. Smad4: gatekeeper gene in head and neck squamous cell carcinoma. J Clin Invest 2009;119:3208–11.
50. Samet JM, Avila-Tang E, Boffetta P, Hannan LM, Olivo-Marston S, Thun MJ, et al. Lung cancer in never smokers: clinical epidemiology and environmental risk factors. Clin Cancer Res 2009;15:5626–45.
51. Devesa SS, Bray F, Vizcaino AP, Parkin DM. International lung cancer trends by histologic type: male:female differences diminishing and adenocarcinoma rates rising. Int J Cancer 2005;117:294–9.