Heredity (2018) 121:537–547
https://doi.org/10.1038/s41437-018-0063-4

the geneticssociety

ARTICLE

Collective effects of common SNPs and risk prediction in lung cancer

Xiaoyun Lei[1] · Dejian Yuan[2] · Zuobin Zhu[3] · Shi Huang[1]

Received: 8 December 2017 / Accepted: 10 January 2018 / Published online: 10 March 2018

Abstract

Lung cancer is the leading cause of cancer deaths in both men and women in the US. While most sporadic lung cancer cases are related to environmental factors such as smoking, genetic susceptibility may also play an important role and a number of lung cancer associated single-nucleotide polymorphisms (SNPs) have been identified although many remain to be found. The collective effects of genome-wide minor alleles of common SNPs, or the minor allele content (MAC) in an individual, have been linked with quantitative variations of complex traits and diseases. Here we studied MAC in lung cancer using previously published SNPs data sets (US and Finland samples) and found higher MAC in cases relative to matched controls. A set of 5400 SNPs with MA (MAF < 0.5) more common in cases ($P < 0.08$) and linkage disequilibrium (LD) $r^2 = 0.3$ was found to have the best predictive accuracy. These results identify higher MAC in lung cancer susceptibility and provide a meaningful genetic method to identify those at risk of lung cancer.

Introduction

Lung cancer is the leading cause of cancer death in both men and women in the U.S and an estimated 158,040 Americans are expected to die from lung cancer in 2015, accounting for approximately 27% of all cancer deaths (CDC, 2014). The most common environmental risk factor for sporadic lung cancer is smoking and radon (Alberg and Samet, 2003). However, there are large variations in an individual's susceptibility to lung cancer and the heritability of lung cancer is estimated to be 8–14% (Czene et al., 2002; Hemminki et al., 2001). Only a fraction of smokers (~15%) will develop lung cancer in their lifetime, and non-smokers also can develop lung cancers (Spitz et al., 2003). A number of cancer genes such as K-ras, p53, Rb, EGFR, HER2-neu have been identified whose mutations contribute to lung cancers (Ding et al., 2008; Iggo et al., 1990; Johnson et al., 2001; Paez et al., 2004; Takahashi et al., 1989).

Efforts to identify quantitative susceptibility loci in lung cancer have mostly involved genome-wide association studies (GWAS) and identified a number of lung cancer risk single-nucleotide polymorphisms (SNPs; Amos et al., 2008; Landi et al., 2009; Ryan et al., 2015; Zhu et al., 2008). However, they account for a very small fraction of lung cancer cases and their mechanisms of action remain largely unknown (Gibson, 2012).

Known predictive models of lung cancers mostly use smoking status, radon exposure, and family history (Spitz et al., 2007). However, these models cannot predict prebirth risk or risk long before incidence. Researchers have also used a set of susceptibility loci to create a genetic risk score to better predict lung cancer risk (Jostins and Barrett, 2011; Li et al., 2012; Weissfeld et al., 2015). But these predictions were generally poor and not meaningful for clinical use. It has been shown that many complex traits or diseases are associated with an accumulation of enormously large numbers of variants of small effects (Boyle et al., 2017; Purcell et al., 2009).

An allele can be a major allele or minor allele (MA) according to its frequency in the population and the minor allele has frequency <0.5. Most known risk alleles are MAs (Park et al., 2011). We have shown that the collective

These authors contributed equally: Xiaoyun Lei, Dejian Yuan.

Electronic supplementary material The online version of this article (https://doi.org/10.1038/s41437-018-0063-4) contains supplementary material, which is available to authorized users.

✉ Shi Huang
huangshi@sklmg.edu.cn

[1] Center for Medical Genetics, School of Life Sciences, Central South University, 110 Xiangya Road, Changsha, Hunan 410078, China

[2] Department of Birth Health and Heredity, Liuzhou Municipal Maternity and Child Healthcare Hospital, Liuzhou 545000, China

[3] Department of Genetics, Xuzhou Medical University, Xuzhou, Jiangsu 221004, China

Table 1 Number of individuals and SNPs in the final (post-QC) data set

Description	US data set		Finland data set		1kGP
	Cases	Controls	Cases	Controls	Controls
Participants	1209	968	1139	860	308
SNPs	511,807	511,807	512,363	512,363	~390,000

The number of cases and controls from each cohort in the final analysis are listed. SNP refers to the number of SNPs that passed QC

effects of a genome-wide collection of MAs in an individual are linked with risk for Parkinson's disease (Zhu et al., 2015), reproductive fitness (Yuan et al., 2014), diabetes (Gui et al., 2017; Lei and Huang, 2017) and schizophrenia (He et al., 2017). The MA content (MAC) of an individual may be at optimal balance with negative selection on both too high or too low MAC values (Yuan et al., 2014). MAC has also been linked with lung cancer in a mouse lung cancer model (Yuan et al., 2014). Lung cancer is a complex disease with causal factors not yet completely identified. Thus, we suspected that MAC may play a role in lung cancer. If a few major effect mutations can cause cancer, it is not unexpected that numerous minor effect mutations or the so called passenger mutations may also increase cancer risk (Gibson, 2012; McFarland et al., 2017).

We here aimed to study the overall level of genome-wide randomness in lung cancer cases relative to controls as measured by total MA amounts in an individual. We also attempted to identify a set of MAs that can predict lung cancer risks.

Materials and methods

SNPs data sets

We downloaded from database of Genotypes and Phenotypes (dbGaP) (https://www.ncbi.nlm.nih.gov/gap) one case control GWAS data set, phs000336.p1.v1. Its dbGaP web page described 5699 cases and 5818 controls, but in fact, only ~3900 controls and ~3800 cases were available to be downloaded, and it consisted of SNPs data sets from five Illumina platforms: (1) HumanHap240Sv1.0 (genotyping ~30 cases and ~1100 controls at 243,991 Oligos/SNPs), (2) HumanHap300v1.1 (genotyping the same individuals as HumanHap240Sv1.0 at 317,503 Oligos/SNPs), (3) HumanHap550v3.0 (genotyping ~770 cases and ~850 controls at 561,466 Oligos/SNPs), (4) Human610_Quadv1_B (genotyping ~3000 cases and ~1800 controls at 620,901 Oligos/SNPs), (5) Human1M-Duov3_B (genotyping ~150 controls at 1,199,187 Oligos/SNPs).

Since the data set 1 above shared only ~330,000 SNPs with other platforms data sets, it was excluded. In addition, identical individuals were also removed, who were genotyped by both Human610_Quadv1_B and HumanHap550v3.0 platforms. The remaining samples (6576 individuals [3782 cases and 2794 controls] with 551,741 SNPs overlapped) were genotyped by Illumina Human610_Quadv1_B or Human1M-Duov3_B or HumanHap550v3.0 platforms. They came from three studies: (1) the Cancer Prevention Study II Nutrition Cohort (CPS-II) (enrolled in the U.S.), (2) the Alpha-Tocopherol, Beta-Carotene Cancer Prevention Study (ATBC) (enrolled in Finland), and (3) the Prostate, Lung, Colon and Ovary Study (PLCO) (enrolled in the U.S.) (1994; Calle et al., 2002; Hayes et al., 2005; Landi et al., 2009). So, these individuals were from US and Finland. Cases were admitted based on chest X ray examination. Participants are all European descendant. We also downloaded from 1000 Genomes Project (1kGP) (http://www.internationalgenome.org) involving 2504 individuals from multiple population groups with a total of ~84.4 million variants (Auton et al., 2015).

Subjects selection

Principal components analysis (PCA) is common in assessing population structure and genetic background. While the chosen thresholds based on PCA to exclude outliers were somewhat arbitrary in common practice, our priority was to include as many samples as possible when no clear genetic substructures could be found as visually judged from the PCA plot. We used the software GCTA to calculate the value of principal components of each sample and figures were plotted with R version 3.2.2. We removed individuals that appeared to be outliers. As illustrated in Supplementary Figure S1, even though all individuals are of European descent, US samples and Finland samples were clustered differently. So we performed separate analyses for these two different sets of samples. For the 3580 US samples, we selected these PC value ranges: $-0.005 < PC1 < 0.015$, $-0.01 < PC2 < 0.005$, and $-0.04 < PC3 < 0.01$ (Supplementary Figure S2A and S2B). For the 2996 Finland samples, we selected these PC value ranges: $-0.02 < PC1 < 0.03$, $-0.03 < PC2 < 0.02$, and $-0.04 < PC3 < 0.04$ (Supplementary Figure S3A and S3B). For 1kGP samples, PCA was also performed (Supplementary Figure S2C and S2D, Supplementary Figure S3C and S3D).

SNPs quality control (QC)

We next performed a SNP-level set of QC steps. SNPs were filtered by removing those with >5% non-informative calls in the population, and those not following the Hardy-Weinberg equilibrium in either the case group or the control group ($P < 0.0001$ chi square test), and those with MAF <

0.01. Only autosome SNPs were used. Samples with >10% missing SNPs and non-founders were excluded (i.e. only parents were retained in cases where their children were also sampled). Overall, these steps resulted in two data sets with ~510,000 SNPs (from 551,741 in phs000336.p1.v1). The description of cleaned up data sets is shown in Table 1.

Statistical analysis

Minor allele frequency (MAF) refers to the frequency at which the second most common allele occurs in a given population. MAs were defined as those alleles with MAF < 0.5 in the control group. The MAC of an individual is the number of MAs divided by the total number of SNPs examined (Yuan et al., 2014). We used a custom script to calculate the MAC values of case and control groups (https://github.com/health1987/dist). Difference in the average MAC value was compared by t test. PLINK was used to calculate a linkage disequilibrium (LD) (r^2) score for each pair of SNPs in a window of 200 kb SNPs, and one SNP from the pair was excluded if $r^2 > 0.4$. To justify this r^2 threshold, we also tested the results at other r^2 levels (i.e. $r^2 = 0.05$, $r^2 = 0.1$, $r^2 = 0.2$, $r^2 = 0.3$, $r^2 = 0.4$, $r^2 = 0.5$, $r^2 = 0.6$, $r^2 = 0.7$, $r^2 = 0.8$ $r^2 = 0.9$ and $r^2 = 1$).

For each GWAS data set (the US data set and the Finland data set), we used PLINK (Purcell et al., 2007) to perform logistic regression test, which allows for multiple covariates when testing for disease trait SNP association, and obtained regression coefficient (beta) and asymptotic P-value for each SNP. A positive regression coefficient (beta) means that the minor allele increases risk mean. Logistic regression details including getting beta of SNPs using PLINK (Purcell et al., 2007) have been reported previously (Hagenaars et al., 2017).

Risk prediction model

Since the US cohort includes 2177 individuals (1209 cases and 968 controls) and was a bit more than the 1999 individuals (1139 cases and 860 controls) in the Finland cohort, so we only performed risk prediction analysis in the former data set. The US data set was randomly separated into training (716 cases/590 controls), validation 1 (242 cases/193 controls), and validation 2 (251 cases/185 controls) cohorts at a ratio of 6:2:2.

Using logistic regression test, we obtained regression coefficient (beta) for each SNP in the GWAS data set that was used as the training set. Since a positive beta means that the minor allele increases risk mean, we used four methods to create genetic risk score (GRS): (1) adding up the weighted value of each risk allele regardless whether the beta was positive or negative, (2) adding up the non-weighted value of each risk allele regardless whether the beta was positive or negative, (3) adding up the weighted value of each risk allele with positive beta, and (4) adding up the non-weighted value of each risk allele with positive beta.

$$\mathrm{wGRS} = \sum_{i=1}^{n} \mathrm{beta_{SNPi}} + 0.5 * \sum_{j=1}^{m} \mathrm{beta_{SNPj}} \tag{1}$$

SNPi represents MAs in homozygous state and SNPj represents MAs in heterozygous state. A custom script was used to calculate the total weighted genetic risk score (wGRS) according to equation (1).

$$\mathrm{GRS} = \sum_{i=1}^{n} \mathrm{SNP_i}, \tag{2}$$

where the GRS was the total number of MAs of SNPs chosen. For each locus, SNP_{i} is 0, 1 or 2 depending on whether the site was homozygous major alleles, heterozygous, or homozygous minor alleles. We obtained GRS of individuals by using a custom script according to equation (2) (Supplementary Materials).

$$\mathrm{wGRS_{positive}} = \sum_{i=1}^{n} \mathrm{beta_{SNPi}} + 0.5 * \sum_{j=1}^{m} \mathrm{beta_{SNPj}} \tag{3}$$

Only SNPs with positive beta were considered for equation (3).

$$\mathrm{GRS_{positive}} = \sum_{i=1}^{n} \mathrm{SNP_i} \tag{4}$$

Only SNPs with positive beta were considered for equation (4).

Based on the logistic regression test, we also obtained asymptotic P-value for each SNP in the GWAS training set. In order to obtain a best model for risk prediction, SNPs at 19 different P-values (<0.001, <0.003, <0.005, <0.007, <0.009, <0.01, <0.02, <0.03, <0.04, <0.05, <0.06, <0.07, <0.08, <0.09, <0.1, <0.3, <0.5, <0.7 and <1) in training data set were chosen at first among all SNPs studied here (i.e. 19 risk prediction models were created). In addition, to avoid overfitting of the prediction model on the training set from which the SNPs set was derived, LD clumping was performed in the training cohort. Different SNPs were chosen to construct the genetic risk score (GRS) at different P-value thresholds (same as above) and different LD r^2 thresholds ($r^2 = 1$, $r^2 = 0.9$, $r^2 = 0.8$, $r^2 = 0.7$, $r^2 = 0.6$, $r^2 = 0.5$, $r^2 = 0.4$, $r^2 = 0.3$, $r^2 = 0.2$, $r^2 = 0.1$ and $r^2 = 0.05$) (i.e, 19 × 11 = 209 models were created). Here, 19 + 209 = 228 models were built. So, the final total number of models was 4 × 228 = 912 models for the US training data set (716 cases/590 controls), taking into account of four methods (wGRS, GRS, $\mathrm{wGRS_{positive}}$ and $\mathrm{GRS_{positive}}$) and 228 models per method.

Risk prediction evaluation

We performed the internal validation twice to estimate the predictive power of the models; in the first phase, a models' performance was evaluated based on validation 1 data set (242 cases/193 controls); only those performing well could enter the second phase in which the models were evaluated based on validation 2 data set (251 cases/185 controls); results from validation 2 were used to quantify model performance. Each experiment's discriminatory capability was evaluated using the receiver operating characteristic (ROC) curve. We then calculated the AUC using GraphPad Prism 6 and the "pROC" R package. In order to obtain a MA set performing well in risk prediction by using one of the four methods (wGRS, GRS, $\text{wGRS}_{\text{positive}}$ and $\text{GRS}_{\text{positive}}$), 228 models were respectively constructed for each method based on *P*-value from logistic regression test and LD clumping r^2 or no LD clumping. We then obtained AUC of each model in the validation 1 data set.

The model performing well and stably in two internal validation experiments was chosen as the final prediction model for each method. The Hosmer–Lemeshow test is a statistical test for goodness of fit for logistic regression models, which is used frequently in risk prediction models (Alba et al., 2017; Hosmer et al., 1997; Krag et al., 1998). We used "RSADBE" R packages and "pscl" R packages to perform the Hosmer–Lemeshow goodness-of-fit test (HL test) for assessing this model calibration; calibration refers to the accuracy of absolute risk estimates (Alba et al., 2017); when the *P*-value from this test is larger than 0.05, the model can be considered as well calibrated.

Comparison with existing methods

Since GRS proposed above is also a sort of polygenic risk score (PRS) (Purcell et al., 2009), assuming the collective effect of many SNPs, we compared its prediction accuracy with other PRS-based methods (such as PRSice (Euesden et al., 2015), LDpred (Vilhjalmsson et al., 2015), and AnnoPred (Hu et al., 2017)). For PRSice (Euesden et al., 2015; Hagenaars et al., 2017), SNPs were first chosen based on passing both LD pruning (0, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1) and GWAS p-value thresholding (0.001, 0.003, 0.005, 0.007, 0.009, 0.01, 0.02, 0.03, 0.04, 0.05, 0.06, 0.07, 0.08, 0.09, 0.1, 0.3, 0.5, 0.7, 1). Polygenic scores were then calculated by summing up alleles associated with lung cancer, weighted by odds ratio from logistic regression test. For LDpred (Vilhjalmsson et al., 2015), SNPs were screened first using different fractions of causal variants (1, 0.3, 0.1, 0.03, 0.01, 0.003, 0.001, 0.0003, 0.0001); the posterior effect size of each SNP was then inferred based on odds ratio and LD information followed by risk score calculations. AnnoPred (Hu et al., 2017) is similar to LDpred but leverages functional annotations to reevaluate SNPs effect. All the evaluation and identification for models were performed in the two phased internal validation as we did for GRS.

Pathway enrichment analysis

We used ANNOVAR (Wang et al., 2010) to annotate the genes associated with the set of risk SNPs identified by the above analysis. We used WebGestalyR (Wang et al., 2013) tool to check the pathways associated with these genes in the Kyoto Encyclopedia of Genes and Genomes database (KEGG). The enriched pathways in the risk SNPs set were compared by chi square test with a group of SNPs chosen randomly.

Risk prediction in other population

For the model performing the best in the US population, its predictive value was also estimated in another independent cohort (the Finland population).

Results

Enrichment of minor alleles in lung cancer cases

We used a previously published GWAS data set of lung cancer case and control cohorts for our studies (Landi et al., 2009). The cleaned data sets after removing genetic outliers were described in Table 1. Total number of samples used here was 2348 cases and 1828 controls including the US cohort and the Finland cohort. In each cohort, we used the control data set for identifying minor allele status, and calculated the MAC value of each individual.

For the US samples who were all of European descent, there were 511,807 autosomal SNPs after QC. For SNPs set with $\text{MAF} < 0.5$ (including 511,547 SNPs), the average MAC value of controls was significantly lower than that of cases (Fig. 1a and Table 2). We further performed the MAC analyses by using subsets of SNPs. After filtering by LD at different r^2 levels (Table 2), we obtained many subsets of autosomal SNPs. For example, when $r^2 = 0.05$, there were 21,974 SNPs, and the average MAC value of controls was also significantly lower than that of cases (Table 2). The results were similar in other subsets of various r^2 levels (Table 2).

In addition, we repeated the MAC analysis by comparing the cases of the US cohort with a different control group, the European group of 1kGP (Auton et al., 2015). The US cohort shared 398,279 with 1kGP. Based on PCA (Supplementary Figure S2C and S2D), 210 unrelated individuals from 1kGP remained after removal of outliers. They were

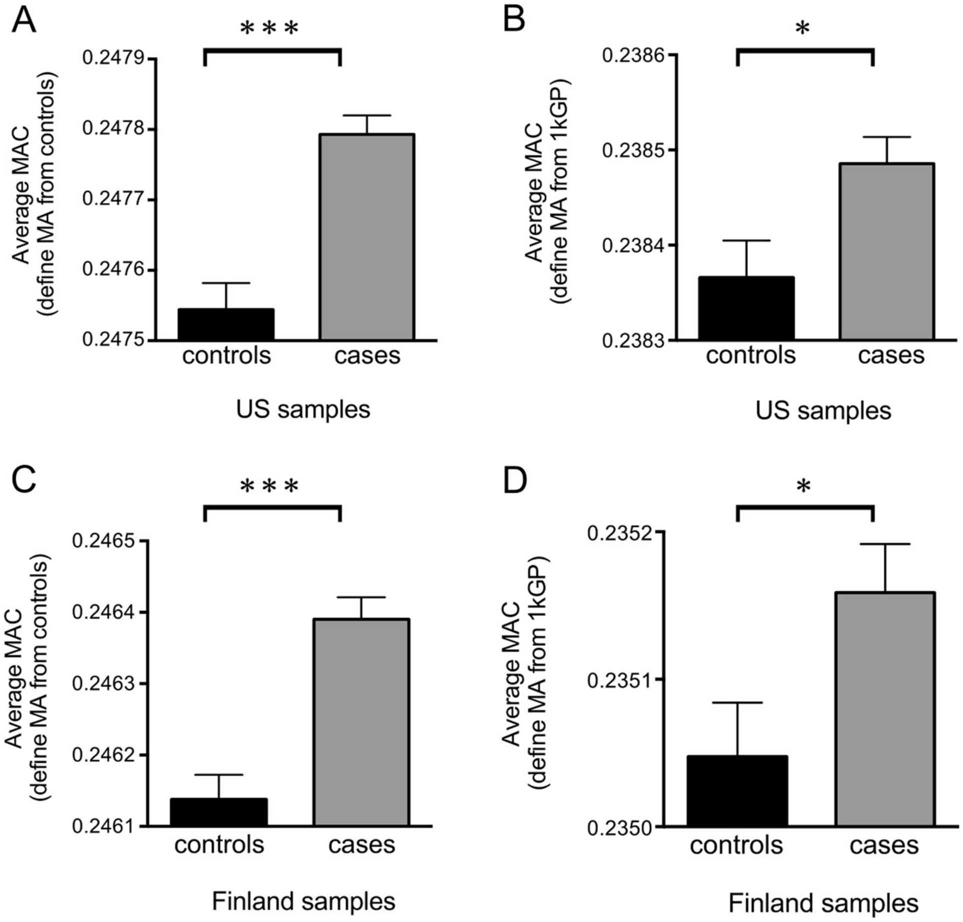

Fig. 1 Comparison of average MAC values. Shown are US samples (**a**, **b**) and Finland samples (**c**, **d**). Either all autosomal SNPs (**a**, **c**) or only those shared with 1 kGP were used for analysis (**b**, **d**). *** $P<0.001$, and * $P<0.05$. Student's *t* test. Standard error of mean (SEM) values are shown

similar to the US group in genetic background and mostly Northern and Western European ancestry (CEU) and Iberian ancestry of Spain (IBS). Among the 398,279 SNPs shared with 1kGP, only 7348 SNPs showed different MA status between the US control group and 1kGP. The majority of these 7348 SNPs had MAFs near 0.5 (~99.8% SNPs with MAF > 0.4), which would make the MA assignment less certain, and were hence excluded. We further removed those SNPs with MAF = 0.5 or MAF = 0 in the US control group or 1kGP group. For the remaining 389,969 SNPs (MAF > 0 and < 0.5 in 1kGP), the average MAC of cases was significantly higher than controls (Fig. 1b and Supplementary Table S1). Therefore, the result of higher MAC in cases could be verified by using a cohort not associated with the original case control studies. For subsets of SNPs based on LD clumping at many r^2 levels (Supplementary Table S1), no difference in MAC was found for some subsets that had lower numbers of SNPs (19,954 SNPs remained at $r^2 = 0.05$ and 42,958 SNPs remained at $r^2 = 0.1$). However, for subsets with relatively more SNPs, the average MAC of cases was again significantly higher than controls (Supplementary Table S1). We also examined SNPs with MAF <0.5 but >0.05, and found the average MAC of cases to be more significantly higher than that of controls (data shown in Supplementary Table S2).

For the Finland samples, there were 512,363 autosomal SNPs after QC. For SNPs set with MAF <0.5 (including 512,106 SNPs), the average MAC value of control was significantly lower than that of cases (Fig. 1c and Table 3). We further performed the MAC analyses on subsets of SNPs based on LD clumping and observed higher MAC in cases in all subsets (Table 3).

We also compared the cases of the Finland cohort with another control group, the European group of 1kGP. The Finland cohort shared 398,138 SNPs with 1kGP. Based on PCA (Supplementary Figure S3C and S3D), 98 unrelated individuals from 1kGP remained after removal of outliers, and were mainly of Finish ancestry (FIN). Among the 398,138 SNPs shared with 1kGP, 10,885 SNPs showed different MA status between the Finland control group and 1kGP. The majority of these 10,885 SNPs had MAFs near 0.5 (~99.9% SNPs with MAF > 0.4), which would make the MA assignment less certain, and were hence deleted. We also removed those SNPs with MAF = 0.5 or MAF = 0 in the Finland control group. For the remaining 385,616 SNPs (MAF > 0 and MAF < 0.5 in 1kGP), the average MAC of

Table 2 MAC (mean ± SD) comparison in the US samples

SNPs set	Number	MAC controls ($n = 968$)	MAC cases ($n = 1209$)
Total loci	511,547	0.24754 ± 1.17E-03	**0.24779 ± 9.35E-04*** **
$r^2 = 0.05$	21,974	0.15483 ± 1.70E-03	**0.15528 ± 1.67E-03*** **
$r^2 = 0.1$	45,471	0.16036 ± 1.28E-03	**0.16073 ± 1.18E-03*** **
$r^2 = 0.2$	88,752	0.17955 ± 1.09E-03	**0.17994 ± 9.15E-04*** **
$r^2 = 0.3$	130,675	0.19701 ± 1.01E-03	**0.19736 ± 8.09E-04*** **
$r^2 = 0.4$	174,501	0.21151 ± 9.97E-04	**0.21182 ± 7.84E-04*** **
$r^2 = 0.5$	222,223	0.22287 ± 1.01E-03	**0.22318 ± 7.88E-04*** **
$r^2 = 0.6$	271,731	0.23141 ± 1.04E-03	**0.23169 ± 7.97E-04*** **
$r^2 = 0.7$	320,982	0.23803 ± 1.05E-03	**0.23828 ± 8.25E-04*** **
$r^2 = 0.8$	366,757	0.24275 ± 1.09E-03	**0.24301 ± 8.47E-04*** **
$r^2 = 0.9$	407,569	0.24551 ± 1.11E-03	**0.24575 ± 8.76E-04*** **
$r^2 = 1$	492,053	0.24894 ± 1.16E-03	**0.24919 ± 9.25E-04*** **

The *t* test was performed to get *P*-values, and boldface indicates significant *P*-values. ***$P < 0.001$, **$P < 0.01$, and *$P < 0.05$

cases was significantly higher than controls (Fig. 1d and Supplementary Table S3). Therefore, the result of higher MAC in cases could be verified by using a cohort not associated with the original case control studies. We further analyzed subsets of SNPs-based LD clumping at many different r^2 levels (Supplementary Table S3). Higher MAC in cases were observed for all SNPs subsets with relatively large number of SNPs (>37,120). We also examined SNPs with MAF < 0.5 but >0.05 (removed rare SNPs), and found the average MAC of cases to be more significantly higher than that of controls (data shown in Supplementary Table S4).

Distinguish cases from controls

Since the MAC of cases was higher than that of controls, we aimed to distinguish cases from controls based on MAC values. However, although the average MAC of cases was significantly higher than controls ($P < 0.0001$), MAC values alone could not produce clear separation of cases from controls (Fig. 2a, b). We therefore generated a wGRS by taking into account of beta values from logistic regression analyses. The total MA number of each individual was then converted into a total wGRS by adding the coefficient of each MA (major alleles were not counted). By converting MAC into the wGRS, the results showed clear separation of cases and controls in both the US and the Finland data sets (Fig. 2c, d).

Risk prediction

We next aimed to obtain a specific set of MAs from a training data set that could be used to predict lung cancer risk for an unrelated data set (validation data set). Since the US data set and Finland data set were genetically different groups (see PCA plot Supplementary Figure S1) and the US data set had larger sample size, we only performed risk prediction studies on the US cohort.

We calculated GRS by counting minor alleles only or by also taking into account regression coefficient (beta). For a MA, when its frequency in the case group is larger than that in the control group, the beta would be positive. We generated four types of GRS metrics. GRS and $\text{GRS}_{\text{positive}}$ were just minor allele counts and $\text{GRS}_{\text{positive}}$ only counted MAs with positive beta. wGRS had MA counts weighted by beta of both positive and negative, and $\text{wGRS}_{\text{positive}}$ only weighted MAs with positive beta. For each score, 228 models were created according to *P*-values from logistic regression test and LD r^2 levels or no LD clumping (Fig. 3). Then, in the first phase internal validation, we used the ROC curve and AUC to examine the discriminatory capability of each model in validation 1 data set (242 cases/193 controls).

For wGRS method in validation 1 data set (Fig. 3a), 30 out of 228 models had AUC ≥ 0.55, and were further evaluated in validation 2 data set (251 cases/185 controls). The model performing the best in the validation 2 data set was identified as LD $r^2 = 0.2$, $P = 0.07$, AUC = 0.554 [95%CI = 0.4996–0.6075]. Similar analyses identified the best model for the other three methods as shown in Table 4.

As shown in Table 4, the model with highest AUC (0.5591, 95% CI:0.5051–0.613) was created by the $\text{wGRS}_{\text{positive}}$ method. The $\text{wGRS}_{\text{positive}}$ model consisted of 5400 SNPs (LD $r^2 = 0.3$ and GWAS *P*-value = 0.08) (Supplementary Table S5). Hosmer–Lemeshow goodness-of-fit test found this best model to be well calibrated ($P = 0.46$). As a comparison, we also similarly analyzed our previous work on Parkinson's disease (Zhu et al., 2015). Among the ~820,000 SNPs analyzed, there were ~420,000 SNPs with positive beta. The risk prediction model

Table 3 MAC (mean ± SD) comparison in the Finland samples

SNPs set	Number	MAC controls ($n = 860$)	MAC cases ($n = 1139$)
Total loci	512,106	0.24614 ± 1.01E-03	**0.24639 ± 1.03E-03*** **
$r^2 = 0.05$	20,103	0.14045 ± 1.89E-03	**0.14106 ± 1.93E-03*** **
$r^2 = 0.1$	43,653	0.15348 ± 1.39E-03	**0.15401 ± 1.48E-03*** **
$r^2 = 0.2$	86,923	0.17506 ± 1.10E-03	**0.17546 ± 1.21E-03*** **
$r^2 = 0.3$	128,705	0.19338 ± 1.01E-03	**0.19374 ± 1.07E-03*** **
$r^2 = 0.4$	172,050	0.20880 ± 9.45E-04	**0.20912 ± 1.01E-03*** **
$r^2 = 0.5$	218,653	0.22030 ± 9.11E-04	**0.22059 ± 9.76E-04*** **
$r^2 = 0.6$	266,563	0.22928 ± 9.05E-04	**0.22957 ± 9.69E-04*** **
$r^2 = 0.7$	314,558	0.23597 ± 9.11E-04	**0.23625 ± 9.63E-04*** **
$r^2 = 0.8$	359,826	0.24097 ± 9.33E-04	**0.24125 ± 9.72E-04*** **
$r^2 = 0.9$	402,640	0.24442 ± 9.52E-04	**0.24468 ± 9.86E-04*** **
$r^2 = 1$	477,767	0.24813 ± 9.90E-04	**0.24838 ± 1.01E-03*** **

The *t* test was performed to get *P*-values, and boldface indicates significant *P*-values. ***$P < 0.001$, **$P < 0.01$, and *$P < 0.05$

performing the best was a wGRS model containing ~37,000 SNPs with MAF < 0.4 and P-value < 0.05. The AUC (0.5795, 95% CI: 0.5391–0.6199) was higher than of that using only SNPs with positive beta (~20,000 SNPs, AUC = 0.555, 95% CI: 0.5149–0.5951) or that with negative beta (~17,000 SNPs, AUC = 0.5713, 95% CI: 0.5314–0.6112). Therefore, the two scoring methods wGRS and wGRS$_{\text{positive}}$ may perform differently in different diseases.

Comparison with existing methods

We used three previously published PRS methods to create risk scores based on the US training data set, PRSice (Euesden et al., 2015), LDpred (Vilhjalmsson et al., 2015) and AnnoPred (Hu et al., 2017). These methods were similarly evaluated in two internal validation analyses, and the best performing models were shown in Table 4. They all appeared to have lower AUC values than the wGRS$_{\text{positive}}$ model.

Pathway enrichment

Using ANNOVAR (Wang et al., 2010), we identified 4832 genes in the best wGRS$_{\text{positive}}$ model containing 5400 risk SNPs. We then used WebGestalyR (Wang et al., 2013) to look for KEGG pathways associated with each of these genes. A total of 39 KEGG pathways were identified with false discovery rate <0.05 (Supplementary Table S6). We also similarly studied a 5400 SNPs set chosen at random, which corresponded to 4954 genes. These genes were enriched in some pathways (Supplementary Table S7). We identified ten pathways that were enriched in the risk set relative to the random SNPs set (Table 5). Some of these pathways are known to be linked to small cell lung cancer, melanoma, prostate cancer (adherens) (Ramteke et al., 2015), and breast cancer (estrogen) (Yamaguchi et al., 2005).

Risk prediction in other population

In addition, for the 5400 SNPs wGRS$_{\text{positive}}$ model performing the best in the US population, its predictive value was low in the Finland population (AUC = 0.4982, 95% CI: 0.4727–0.5238). Thus, the prediction model identified here was highly population specific.

Discussion

In the present study, we showed enrichment of MAs in lung cancer cases relative to matched controls, suggesting a role for the collective effects of polygenic variations in the risk for lung cancer. We also calculated wGRS$_{\text{positive}}$ of each subject based on MA status of SNPs and did risk prediction. We identified a set of MA of common SNPs that can be used to identify subjects at risk of lung cancer.

The result of higher MAC in lung cancer cases is a novel finding not expected by known works on human lung cancers. It confirms the previous result showing MAC association with lung cancer in a mouse lung cancer model (Yuan et al., 2014). Published lung cancer risk SNPs are relatively few in numbers. Therefore, even if these known risk alleles are mostly minor alleles, it may not predict that cases should have more MAs when a genome-wide collection of ~500k SNPs are considered. If most MAs are not related to lung cancer except those few published lung cancer alleles, the average MAC of cases should not be significantly different from the controls.

Our study here further strengthened the observation that human genetic diversities are presently at optimum level

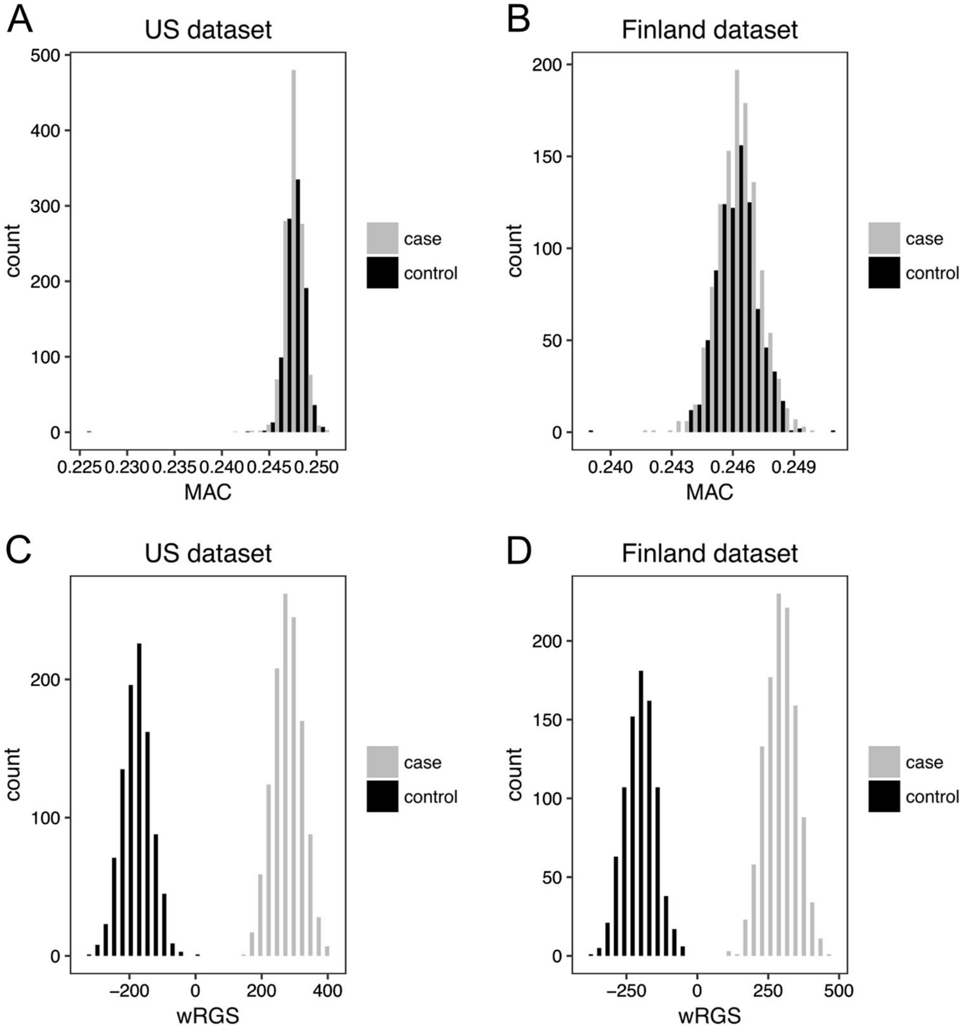

Fig. 2 The distribution pattern of MAC and wGRS. MAC (**a** and **b**) and wGRS (**c** and **d**) values in the US data set (**a** and **c**) and the Finland data set (**b** and **d**) were plotted against the number of individuals

(Huang, 2008; Huang, 2009; Huang, 2016; Yuan et al., 2017; Zhu et al., 2015). While it may only take one or a few major effect errors to cause diseases, it would require the collective effects of many minor effect errors to achieve a similar outcome. Cancer is known to be a disease of random mutations. Individuals with too many inherited random mutations or MAs may need fewer somatic mutations to pass the cancer threshold and hence have high susceptibility to cancer.

AUC has been used in many studies for gauging performance of prediction model (Alba et al., 2017; Kang et al., 2011; O'Connell et al., 2016). Our predictive model of lung cancer was comparable to previous results as indicated by AUC values (Li et al., 2012). Our best predictor model has a AUC of 0.559 as verified by validation experiment. It seems to be low but may still be meaningful. In addition, calibration is one of the most important metrics for prediction models (Alba et al., 2017). Our best predictor model is well calibrated while many previous models did not take this into consideration (Hagenaars et al., 2017; Kang et al., 2011; Lei and Huang, 2017; Li et al., 2012).

The results here indicate interesting differences in the role of MAC between lung cancer and Parkinson's disease (Zhu et al., 2015). Some epidemiological work showed that cancer seemed to occur less frequently in the context of Parkinson's disease (Devine et al., 2011). Only SNPs with positive beta or higher frequency in cases were found useful in prediction models in the case of lung cancer but not Parkinson's disease. Evidence suggests that there is an optimal balance in MAC of an individual (Yuan et al., 2014). Minor alleles are in general under more negative selection but also essential for certain physiological functions such as immunity. Certain diseases may be linked to collective effects of minor alleles with increased frequency in cases, while certain other diseases may also involve a fraction of minor alleles with decreased frequency in cases. As minor alleles are beneficial for adaptive immunity (Yuan et al., 2014), one may speculate that decreased immunity or

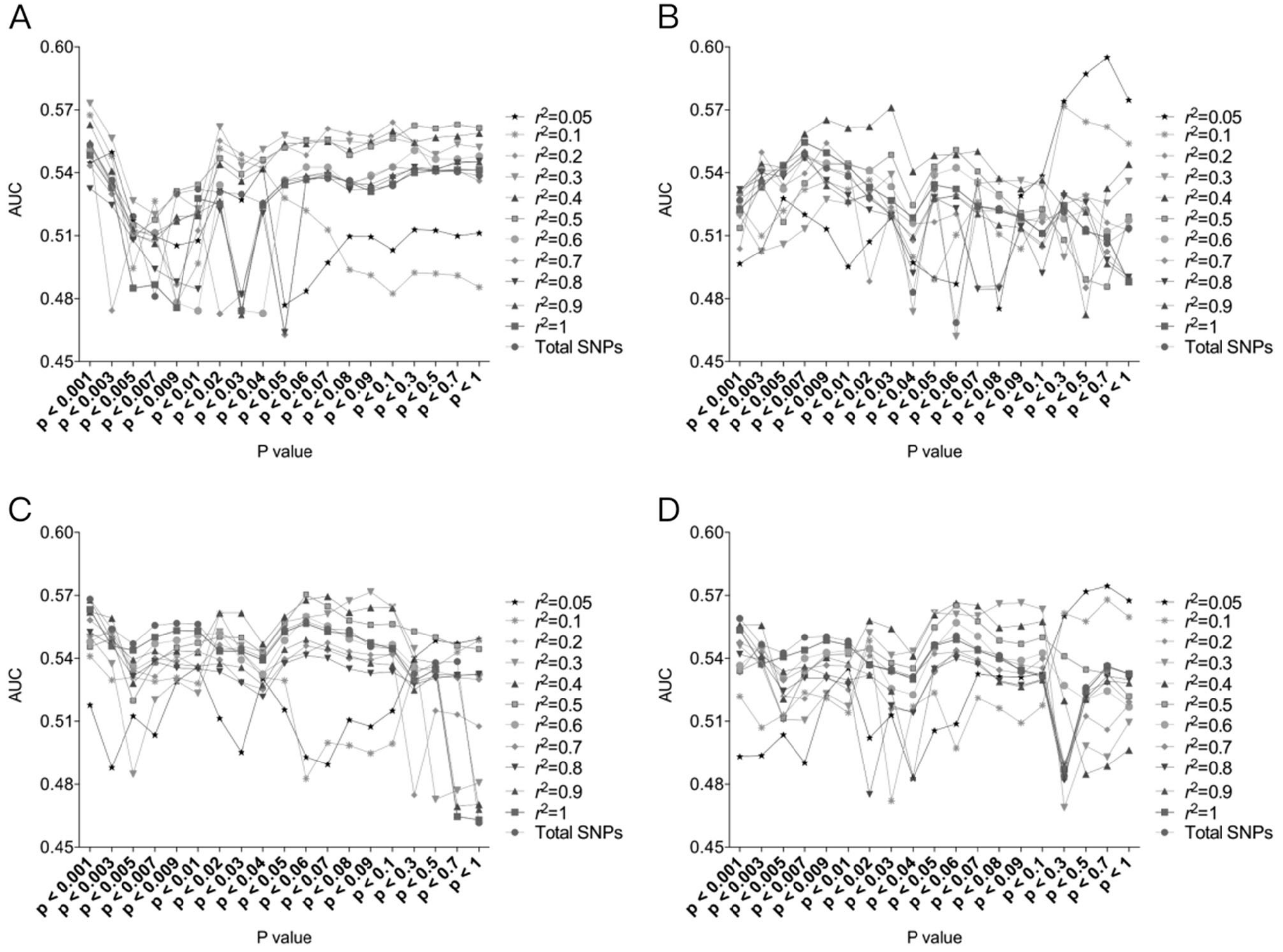

Fig. 3 Discriminatory ability of prediction models. Four different scoring methods are shown. **a** wGRS method; **b** GRS method; **c** $\text{wGRS}_{\text{positive}}$ method and **d** $\text{GRS}_{\text{positive}}$ method

Table 4 AUC of different methods

Methods	AUC	
	Values	95% CI
wGRS	0.554	0.4996–0.6075
GRS	0.5295	0.4747–0.5843
$\text{wGRS}_{\text{positive}}$	**0.5591**	0.5051–0.613
$\text{GRS}_{\text{positive}}$	0.5576	0.5036–0.6116
PRSice	0.5492	0.4952–0.6032
LDpred	0.525	0.4707–0.5794
AnnoPred	0.5226	0.4677–0.5774

Results from validation 2 data set

CI confidence interval

The largest AUC value among different methods is highlighted in boldface

some other physiological functions may play a relatively more important role in Parkinson's disease.

After comparing prediction accuracy of the present $\text{wGRS}_{\text{positive}}$ method with that of previous PRS method, we observed slightly improved results ($\text{wGRS}_{\text{positive}}$: 0.5591 [95% CI 0.5051–0.613]; PRSice: 0.5492 [95% CI 0.4952–0.6032]; LDpred: 0.525 [95% CI 0.4707–0.5794]; AnnoPred: 0.5226 [95% CI 0.4677–0.5774]). That these methods showed similar performance may not be unexpected given that all are based on the theory of polygenic inheritance for complex diseases. However, the PRSice method excludes SNPs from transversion mutations, which may decrease its power (Euesden et al., 2015; Lei and Huang, 2017). In addition, we noticed $\text{GRS}_{\text{positive}}$ method (AUC: 0.5576 [95% CI 0.5036–0.6116]) showed similar results as $\text{wGRS}_{\text{positive}}$ method (AUC: 0.5591 [95% CI 0.5051–0.613]). So, wGRS method only performed slightly better than non-wGRS method. However, since the sample size in our study was relatively small, it remains to be seen how these various risk scoring methods may differ in future studies involving larger sample sizes.

We found the predictive power of our model was population specific (US data set: AUC = 0.5591 [95% CI 0.5051–0.613]; Finland data set: AUC = 0.4982 [95% CI 0.4727–0.5238]). The model was created by using US

Table 5 Pathways enrichment

Pathways in KEGG	Genes of 5400 SNPs for risk prediction	Genes of 5400 SNPs chosen at random	*P*-value, χ^2 test
Neuroactive ligand–receptor interaction	78 (78/4832)	0 (0/4954)	<2.2e-16
Dopaminergic synapse	41 (41/4832)	0 (0/4954)	2.29E-10
Adherens junction	25 (25/4832)	0 (0/4954)	1.12E-06
Cocaine addiction	18 (18/4832)	0 (0/4954)	4.83E-05
Amebiasis	30 (30/4832)	0 (0/4954)	7.80E-08
Serotonergic synapse	33 (33/4832)	0 (0/4954)	1.59E-08
Small cell lung cancer	26 (26/4832)	0 (0/4954)	6.58E-07
Estrogen signaling pathway	29 (29/4832)	0 (0/4954)	1.33E-07
Pancreatic secretion	28 (28/4832)	0 (0/4954)	2.26E-07
Melanoma	22 (22/4832)	0 (0/4954)	5.60E-06

The cancer-related pathways are highlighted in boldface

samples and hence should only work for US samples. This is to be expected since different human populations are known to show group specific SNP profiles (Lei and Huang, 2017).

The 5400 SNPs in our lung cancer prediction model were enriched in small cell lung cancer, melanoma, adherens junction and estrogen signaling pathways. In contrast, randomly chosen SNPs of the same number did not have the same pathway enrichment. Most of these pathways are known to play roles in cancer (Ramteke et al., 2015; Yamaguchi et al., 2005). Our results provide additional evidence for the role of these pathways in lung cancer and may help understand their mechanisms of action in lung cancer.

Acknowledgements This work was supported by the Fundamental Research Funds for the Central Universities of Central South University, the National Natural Science Foundation of China (Grant No. 81171880), and the National Basic Research Program of China (Grant No. 2011CB51001). The data sets used for the analyses described in this manuscript were obtained from dbGaP at http://www.ncbi.nlm.nih.gov/sites/entrez?Db=gap and 1kGP at http://www.1000genomes.org. The dbGaP accession number includes phs000336.p1.v1. We thank the Contributing Investigator (phs000336.p1.v1,Maria Teresa Landi, Neil E. Caporaso, Michael Thun, etc.). The content is solely the responsibility of the authors and does not necessarily represent the official views of the funding agencies.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

References

The ATBC Cancer Prevention Study Group (1994) The alpha-tocopherol, beta-carotene lung cancer prevention study: design, methods, participant characteristics, and compliance. The ATBC Cancer Prevention Study Group. Ann Epidemiol 4(1):1–10

Alba AC, Agoritsas T, Walsh M, Hanna S, Iorio A, Devereaux PJ et al. (2017) Discrimination and calibration of clinical prediction models: users' guides to the medical literature. Jama 318 (14):1377–1384

Alberg AJ, Samet JM (2003) Epidemiology of lung cancer. Chest J 123(1_suppl):21S–49S

Amos CI, Wu X, Broderick P, Gorlov IP, Gu J, Eisen T et al. (2008) Genome-wide association scan of tag SNPs identifies a susceptibility locus for lung cancer at 15q25. 1. Nat Genet 40 (5):616–622

Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO et al. (2015) A global reference for human genetic variation. Nature 526(7571):68–74

Boyle EA, Li YI, Pritchard JK (2017) An expanded view of complex traits: from polygenic to omnigenic. Cell 169(7):1177–1186

Calle EE, Rodriguez C, Jacobs EJ, Almon ML, Chao A, McCullough ML et al. (2002) The American Cancer Society Cancer Prevention Study II Nutrition Cohort: rationale, study design, and baseline characteristics. Cancer 94(9):2490–2501

CDC (2014) Centers for Disease Control and Prevention (CDC). National Center for Health Statistics. CDC WONDER Online Database, compiled from Compressed Mortality File 1999–2012 Series 20 No. 2R.

Czene K, Lichtenstein P, Hemminki K (2002) Environmental and heritable causes of cancer among 9.6 million individuals in the Swedish family-cancer database. Int J Cancer 99(2):260–266

Devine MJ, Plun-Favreau H, Wood NW (2011) Parkinson's disease and cancer: two wars, one front. Nat Rev câncer 11(11):812–823

Ding L, Getz G, Wheeler DA, Mardis ER, McLellan MD, Cibulskis K et al. (2008) Somatic mutations affect key pathways in lung adenocarcinoma. Nature 455(7216):1069–1075

Euesden J, Lewis CM, O'Reilly PF (2015) PRSice: Polygenic Risk Score software. Bioinformatics 31(9):1466–1468

Gibson G (2012) Rare and common variants: twenty arguments. Nat Rev Genet 13(2):135–145

Gui Y, Lei X, Huang S (2017) Collective effects of common SNPs and genetic risk prediction in type 1 diabetes. Clini Genet https://doi.org/10.1111/cge.13193

Hagenaars SP, Hill WD, Harris SE, Ritchie SJ, Davies G, Liewald DC et al. (2017) Genetic prediction of male pattern baldness. PLoS Genet 13(2):e1006594

Hayes RB, Sigurdson A, Moore L, Peters U, Huang WY, Pinsky P et al. (2005) Methods for etiologic and early marker investigations in the PLCO trial. Mutat Res 592(1-2):147–154

He P, Lei X, Yuan D, Zhu Z, Huang S (2017) Accumulation of minor alleles and risk prediction in schizophrenia. Sci Rep 7(1):11661

Hemminki K, Lönnstedt I, Vaittinen P, Lichtenstein P (2001) Estimation of genetic and environmental components in colorectal and lung cancer and melanoma. Genet Epidemiol 20(1):107–116

Hosmer DW, Hosmer T, Le Cessie S, Lemeshow S (1997) A comparison of goodness-of-fit tests for the logistic regression model. Stat Med 16(9):965–980

Hu Y, Lu Q, Powles R, Yao X, Yang C, Fang F et al. (2017) Leveraging functional annotations in genetic risk prediction for human complex diseases. PLoS Comput Biol 13(6):e1005589

Huang S (2008) The genetic equidistance result of molecular evolution is independent of mutation rates. J Comput Sci Syst Biol 1:92

Huang S (2009). Inverse relationship between genetic diversity and epigenetic complexity. http://precedings.nature.com/documents/1751/version/2

Huang S (2016) New thoughts on an old riddle: what determines genetic diversity within and between species? Genomics 108 (1):3–10

Iggo R, Bartek J, Lane D, Gatter K, Harris AL (1990) Increased expression of mutant forms of p53 oncogene in primary lung cancer. Lancet 335(8691):675–679

Johnson L, Mercer K, Greenbaum D, Bronson RT, Crowley D, Tuveson DA et al. (2001) Somatic activation of the K-ras oncogene causes early onset lung cancer in mice. Nature 410 (6832):1111–1116

Jostins L, Barrett JC (2011) Genetic risk prediction in complex disease. Hum Mol Genet 20(R2):R182–R188

Kang J, Kugathasan S, Georges M, Zhao H, Cho JH (2011) Improved risk prediction for Crohn's disease with a multi-locus approach. Hum Mol Genet 20(12):2435–2442

Krag D, Weaver D, Ashikaga T, Moffat F, Klimberg VS, Shriver C et al. (1998) The sentinel node in breast cancer--a multicenter validation study. N Engl J Med 339(14):941–946

Landi MT, Chatterjee N, Yu K, Goldin LR, Goldstein AM, Rotunno M et al. (2009) A genome-wide association study of lung cancer identifies a region of chromosome 5p15 associated with risk for adenocarcinoma. Am J Human Genet 85(5):679–691

Lei X, Huang S (2017) Enrichment of minor allele of SNPs and genetic prediction of type 2 diabetes risk in British population. PLoS One 12(11):e0187644

Li H, Yang L, Zhao X, Wang J, Qian J, Chen H et al. (2012) Prediction of lung cancer risk in a Chinese population using a multifactorial genetic model. BMC Med Genet 13(1):118

McFarland CD, Yaglom JA, Wojtkowiak JW, Scott JG, Morse DL, Sherman MY et al. (2017) The damaging effect of passenger mutations on cancer progression. Cancer Res 77(18):4763–4772

O'Connell PJ, Zhang W, Menon MC, Yi Z, Schroppel B, Gallon L et al. (2016) Biopsy transcriptome expression profiling to identify kidney transplants at risk of chronic injury: a multicentre, prospective study. Lancet 388(10048):983–993

Paez JG, Jänne PA, Lee JC, Tracy S, Greulich H, Gabriel S et al. (2004) EGFR mutations in lung cancer: correlation with clinical response to gefitinib therapy. Science 304(5676):1497–1500

Park JH, Gail MH, Weinberg CR, Carroll RJ, Chung CC, Wang Z et al. (2011) Distribution of allele frequencies and effect sizes and their interrelationships for common genetic susceptibility variants. Proc Natl Acad Sci USA 108(44):18026–18031

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D et al. (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Human Genet 81 (3):559–575

Purcell SM, Wray NR, Stone JL, Visscher PM, O'Donovan MC, Sullivan PF et al. (2009) Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. Nature 460 (7256):748–752

Ramteke A, Ting H, Agarwal C, Mateen S, Somasagara R, Hussain A et al. (2015) Exosomes secreted under hypoxia enhance invasiveness and stemness of prostate cancer cells by targeting adherens junction molecules. Mol Carcinog 54(7):554–565

Ryan BM, Robles AI, McClary AC, Haznadar M, Bowman ED, Pine SR et al. (2015) Identification of a functional SNP in the 3′ UTR of CXCR2 that is associated with reduced risk of lung cancer. Cancer Res 75(3):566–575

Spitz MR, Hong WK, Amos CI, Wu X, Schabath MB, Dong Q et al. (2007) A risk model for prediction of lung cancer. J Natl Cancer Inst 99(9):715–726

Spitz MR, Wei Q, Dong Q, Amos CI, Wu X (2003) Genetic susceptibility to lung cancer: the role of DNA damage and repair. Cancer Epidemiol, Biomark & Prev 12(8):689–698

Takahashi T, Nau MM, Chiba I, Birrer MJ, Rosenberg RK (1989) p53: a frequent target for genetic abnormalities in lung cancer. Science 246(4929):491

Vilhjalmsson BJ, Yang J, Finucane HK, Gusev A, Lindstrom S, Ripke S et al. (2015) Modeling linkage disequilibrium increases accuracy of polygenic risk scores. Am J Hum Genet 97(4):576–592

Wang J, Duncan D, Shi Z, Zhang B (2013) WEB-based GEne SeT AnaLysis Toolkit (WebGestalt): update 2013. Nucleic Acids Res 41(Web Server issue):W77–W83

Wang K, Li M, Hakonarson H (2010) ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res 38(16):e164–e164

Weissfeld JL, Lin Y, Lin H-M, Kurland BF, Wilson DO, Fuhrman CR et al. (2015) Lung cancer risk prediction using common SNPs located in GWAS-identified susceptibility regions. J Thorac Oncol 10(11):1538–1545

Yamaguchi Y, Takei H, Suemasu K, Kobayashi Y, Kurosumi M, Harada N et al. (2005) Tumor-stromal interaction through the estrogen-signaling pathway in human breast cancer. Cancer Res 65(11):4653–4662

Yuan D, Lei X, Gui Y, Zhu Z, Wang D, Yu J et al. (2017) Modern human origins: multiregional evolution of autosomes and East Asia origin of Y and mtDNA. bioRxiv: 101410; https://doi.org/10.1101/101410

Yuan D, Zhu Z, Tan X, Liang J, Zeng C, Zhang J et al. (2014) Scoring the collective effects of SNPs: association of minor alleles with complex traits in model organisms. Sci China Life Sci 57 (9):876–888

Zhu Y, Hoffman A, Wu X, Zhang H, Zhang Y, Leaderer D et al. (2008) Correlating observed odds ratios from lung cancer case–control studies to SNP functional scores predicted by bioinformatic tools. Mutat Res/Fundam Mol Mech Mutagen 639 (1):80–88

Zhu Z, Yuan D, Luo D, Lu X, Huang S (2015) Enrichment of minor alleles of common SNPs and improved risk prediction for Parkinson's disease. Plos One 10(7):e0133421