Polygenic Risk Scores for Prediction of Gastric Cancer Based on Bioinformatics Screening and Validation of Functional lncRNA SNPs

Fujiao Duan, PhD[1,2], Chunhua Song, PhD[2,3], Peng Wang, PhD[2,3], Hua Ye, PhD[2,3], Liping Dai, PhD[2,3], Jianying Zhang, PhD[2,3] and Kaijuan Wang, PhD[2,3]

INTRODUCTION: **Single-nucleotide polymorphisms (SNPs) are used to stratify the risk of gastric cancer. However, no study included gastric cancer–related long noncoding RNA (lncRNA) SNPs into the risk model for evaluation. This study aimed to replicate the associations of 21 lncRNA SNPs and to construct an individual risk prediction model for gastric cancer.**

METHODS: **The bioinformatics method was used to screen gastric cancer–related lncRNA functional SNPs and verified in population. Gastric cancer risk prediction models were constructed using verified SNPs based on polygenic risk scores (PRSs).**

RESULTS: **Twenty-one SNPs were screened, and the multivariate unconditional logistic regression analysis showed that 14 lncRNA SNPs were significantly associated with gastric cancer. In the distribution of genetic risk score in cases and controls, the mean value of PRS in cases was higher than that in controls. Approximately 20.1% of the cases was caused by genetic variation ($P = 1.9 \times 10^{-34}$) in optimal PRS model. The individual risk of gastric cancer in the lowest 10% of PRS was 82.1% (95% confidence interval [CI]: 0.102, 0.314) lower than that of the general population. The risk of gastric cancer in the highest 10% of PRS was 5.75-fold that of the general population (95% CI: 3.09, 10.70). The introduction of family history of tumor (area under the curve, 95% CI: 0.752, 0.69–0.814) and *Helicobacter pylori* infection (area under the curve, 95% CI: 0.773, 0.702–0.843) on the basis of PRS could significantly improve the recognition ability of the model.**

DISCUSSION: **PRSs based on lncRNA SNPs could identify individuals with high risk of gastric cancer and combined with risk factors could improve the stratification.**

SUPPLEMENTARY MATERIAL accompanies this paper at http://links.lww.com/CTG/A721

Clinical and Translational Gastroenterology 2021;12:e00430. https://doi.org/10.14309/ctg.0000000000000430

INTRODUCTION

Gastric cancer is a highly lethal malignancy worldwide, being the fourth most common cancer and the second leading cause of cancer death (1). It is concerned worldwide that Eastern Asian had the highest estimated morbidity and mortality rates (2). Although surgical techniques, radiotherapy, and chemotherapy regimens have helped reduce the incidence and mortality rates of gastric cancer, the overall 5-year survival rate is still only approximately 25% (3,4).

Research in the past 2 decades has revealed that long noncoding RNAs (lncRNAs) with different regulatory functions are effectively fed back into the larger RNA communication network and ultimately regulate the basic protein effectors of cell functions (5,6). The lncRNAs regulate gene expression through a variety of mechanisms, including epigenetic regulation through chromatin remodeling, transcriptional activation or inhibition, mRNA post-transcriptional modification or protein activity regulation (5).

Genetic risk scores (GRSs) can screen out high-risk individuals by evaluating the susceptibility of tumor genes and perform etiological prevention and clinical intervention in advance. However, using the previously developed weighted GRSs to predict cancer risk has some limitations. Researchers are studying that cancer is affected by 1 or more genetic changes, which are often combined with environmental

[1]Medical Research Office, Affiliated Cancer Hospital of Zhengzhou University, Zhengzhou, Henan, China; [2]Key Laboratory of Tumor Epidemiology of Henan Province, Zhengzhou, Henan Province, China; [3]College of Public Health, Zhengzhou University, Zhengzhou, Henan Province, China.
Correspondence: Fujiao Duan, PhD. E-mail: fjduan@126.com. Kaijuan Wang, PhD. E-mail: kjwang@163.com.
Received March 27, 2021; accepted September 23, 2021; published online November 18, 2021

factors to understand the role of genetics in diseases in different populations. Polygenic risk score (PRS) is one way by which people can learn about their risk of developing a disease, based on the total number of changes related to the disease (7,8).

As the biological significance of lncRNA has attracted more and more attention, many efforts have been made to solve the role of lncRNA in cancer. This leads to a large number of studies on the relationship between lncRNA status in gastric cancer and clinical results. Nevertheless, lncRNA single-nucleotide polymorphisms (SNPs) are not involved in the current construction of cancer-related risk models. In this study, we discussed the screening of gastric cancer–related lncRNAs and corresponding functional SNPs by bioinformatics methods and verified the associations with gastric cancer through the population. Based on the results of correlation validation, PRS was used to construct the risk prediction model and explore the optimal model for gastric cancer.

MATERIALS AND METHODS

The flowchart is shown in Figure 1. This study was approved by the ethics committee of Zhengzhou University. All participants were informed and signed written informed consent.

lncRNAs selection

This study integrated 7 online databases (GENCODE.v24, lncRNAdb.v2.0, LNCIpedia.v3.1, Ensembl, CCDS.v18, NONCODE2016, and refse) to overcome the lack of integrity of lncRNAs in a single database. Avalanche workbench 2.0.9 was used to remove the redundancy, and lncRNAs less than 200 bp were eliminated to establish protein data bank and lncRNAs database after redundancy removal.

The nucleic acid sequence alignment algorithm (BlastN) was used to map the probe sequences in the CDF format file corresponding to the Arraystar Human LncRNA microarray V2.0 chip platform to the established protein data bank and lncRNAs databases to construct the chip probe mapping.

We searched and downloaded the microarray data related to gastric cancer in Chinese population (GSE50710, GSE53137, and GSE58828) in the Gene Expression Omnibus database. According to the analysis results of the 3 chips, the intersection was obtained. The multiple of difference was >2.0 and $P < 0.05$, and the differentially expressed lncRNAs were screened.

The lncRNASNP2 database (http://bioinfo.life.hust.edu.cn/lncRNASNP#!/) and the online database RNAfold (http://rna.tbi.univie.ac.at/cgi-bin/RNAWebSuite/RNAfold.cgi) were used to predict the potential biological functions of the differentially expressed lncRNA SNPs and screen out the SNPs that affect the secondary structure of lncRNAs and the binding of microRNAs. Finally, 21 lncRNA SNPs were selected.

Study subjects

This study adopted a case-control study design of 1:1 frequency matching by age (±2 years) and sex and included a total of 1,088 gastric cancer cases and healthy controls.

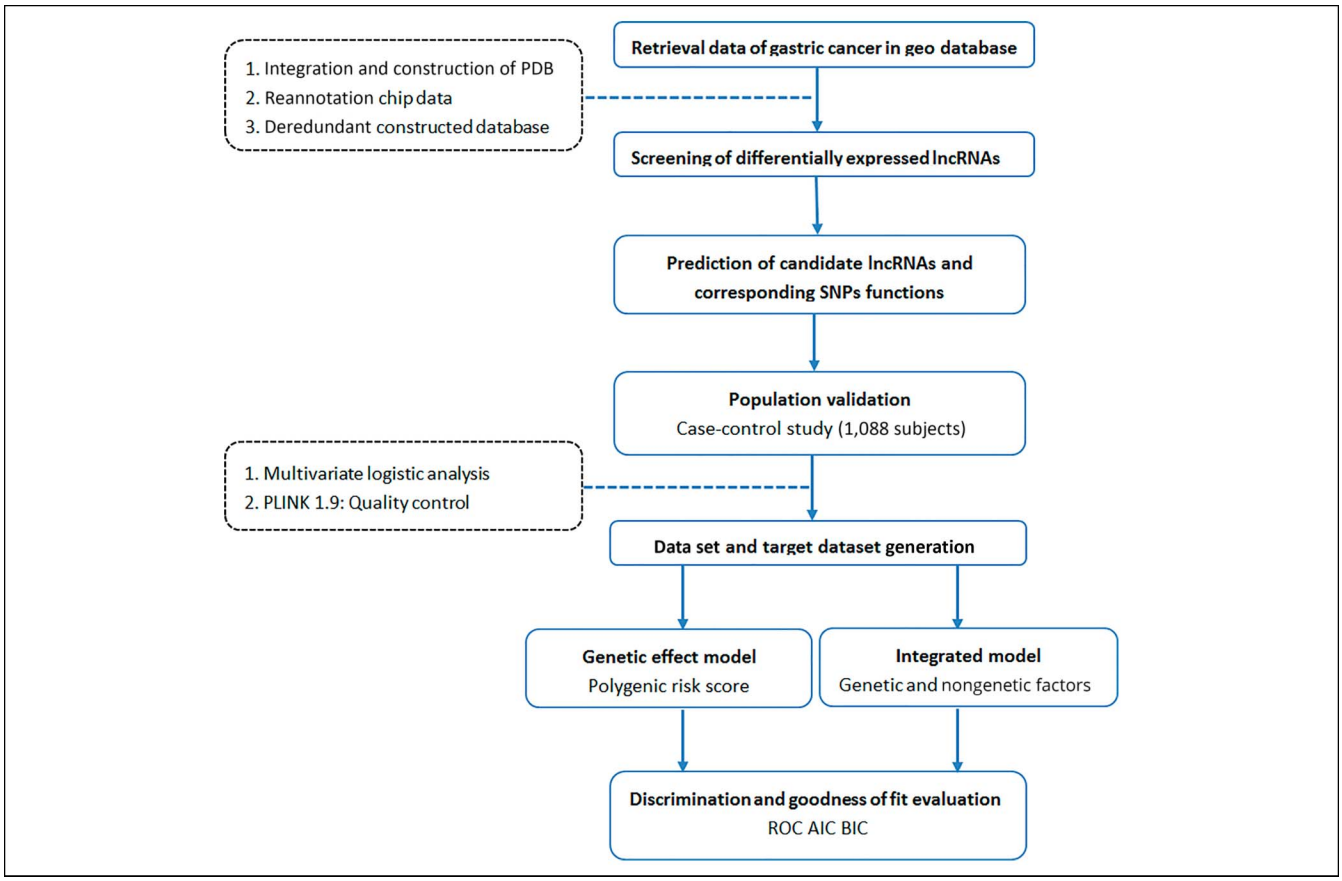

Figure 1. Flowchart of study design. AIC, Akaike information criterion; BIC, Bayesian information criterion; lncRNAs, long noncoding RNAs; PDB, protein data bank; ROC, Receiver operating characteristic; SNP, single-nucleotide polymorphism.

From December 2012 to December 2015, a newly diagnosed gastric cancer patient was confirmed by gastroscopy and postoperative histopathological examination. The selection of gastric cancer was based on the guideline proposed by the Union for International Cancer Control. The inclusion criteria were as follows: (i) first time diagnosis; (ii) those who did not receive radiotherapy, chemotherapy, surgery, and other treatments; (iii) no history of malignancy; and (iv) those with complete clinical, pathological, or follow-up data. The exclusion criteria were as follows: (i) combined with other clinical disorders besides gastric cancer; (ii) cardia cancer; (iii) those who received any preoperative antitumor therapy; and (iv) those with complete clinical, pathological, or follow-up data. The controls were derived from a community population at the same period of the chronic disease epidemiological investigation site in Henan Province. There was no blood relationship with the case, no chronic digestive system disease, and no history of tumors.

Genotyping and data set generation

Based on frequency-matched (1:1) case-control study design to match subjects according to sex and age (±2 years), the blood samples were collected from 544 patients with gastric cancer and 544 normal controls from community. Polymerase chain reaction restriction fragment length polymorphism and created restriction site-polymerase chain reaction restriction fragment length polymorphism were used to genotype SNPs corresponding to lncRNAs (see Supplementary Table S1, Supplementary Digital Content 1, http://links.lww.com/CTG/A721).

Plink 1.9 was used for quality control of related SNPs, association analysis of allele and generation of PRSice-2 (Polygenic Risk Score software) basic data set and target data set.

Risk prediction model

We used bioinformatics screening and association validation SNPs to construct gastric cancer risk prediction model based on PRS. The lncRNA SNPs were put into the prediction model as independent data sets of risk factors.

The PRS was derived from lncRNA SNPs that have been shown to be associated with gastric cancer risk. The effect size (odds ratio [OR]) of PRS was constructed by summing the risk and allele counts (i.e., the subjects had 0, 1, or 2 risk alleles), which was determined by the natural logarithm transformation of risk (i.e., OR (ln)), which was extracted from the results of multiple unconditional logistic regression model. For each participant, we summed the weighted risk allele counts, divided the total number of loci by the average weighted score, and used the average weighted score as a reference.

$$PRSj = \Sigma_j^i n_{ij} \ln(OR_i),$$

where j is the number of SNPs included in the model; n_{ij} is the number of the i-th risk allele (0, 1, or 2); and OR_i is the associated risk value between the risk allele of the i-th SNP and gastric cancer.

Statistical analysis

The SAS software, version 9.2 (SAS Institute, Cary, NC), R-software, version 3.6.1 (the R foundation for statistical computing, Vienna, Austria), and Plink 1.9 (NIH-NIDDK's Laboratory of Biological Modeling, Harvard University, Cambridge, MA) were used for statistical analysis, and all *P* values for statistical significance were 2-sided.

The quantitative variables were expressed by mean ± SD, and the differences between the control group and the case group were analyzed by *t* test; the classification variables were expressed by frequency, and the distribution differences between groups were analyzed by the χ^2 test.

The goodness of fit χ^2 test was used to verify the Hardy-Weinberg equilibrium. The association between lncRNA SNPs and gastric cancer susceptibility was evaluated by using adjusted multivariate unconditional logistic regression model. The PRS summarized the combined effect of the SNPs

Table 1. Basic characteristics of individuals in case and control groups

Variables	Case, n (%) N = 544	Controls (%) N = 544	*t*/χ^2	*P*
Age (mean ± SD)	57.80 ± 12.06	57.02 ± 11.97	1.072	0.284
Sex			0.406	0.524
Men	408 (71.97)	417 (71.97)		
Women	136 (28.03)	127 (28.03)		
Smoking status			4.779	0.029
Nonsmoker	239 (60.30)	275 (69.70)		
Yes	305 (56.07)	269 (30.30)		
Drinking status			0.100	0.752
Nondrinker	348 (69.39)	353(76.82)		
Drinker	196 (30.61)	191 (23.18)		
Family history of tumor			30.990	<0.001
No	413 (88.07)	483 (97.88)		
Yes	131 (11.93)	61 (2.12)		
Helicobacter pylori infection			6.101	0.014
No	82(37.08)	83(50.54)		
Yes	143(63.55)	87(49.46)		

Table 2. Association between the lncRNA SNPs and risk of gastric cancer

SNP(rs#)	lncRNA	Chr./position	RA/ Ref.	OR (95% CI)				
				Per-allele	Heterozygous	Homozygous	Dominant model	Recessive model
rs1859168	lnc-EVX1-3:3	Chr7:27242359	C/A	1.089 (0.920, 1.290)	0.389 (0.275, 0.496)	1.051 (0.769, 1.437)	0.649 (0.492, 0.857)	1.789 (1.386, 2.310)
rs3815254	lnc-MACC1-1:7	Chr 7: 19983014	A/G	0.984 (0.828, 1.171)	1.012 (0.778, 1.316)	0.929 (0.633, 1.364)	0.993 (0.774, 1.275)	0.923 (0.647, 1.315)
rs4784659	lnc-AMFR-1:1	Chr 16: 56387000	C/T	0.554 (0.438, 0.701)	0.420 (0.313, 0.565)	0.572 (0.294, 1.114)	0.438 (0.331, 0.579)	0.710 (0.367, 1.375)
rs579501	lnc-ZNF33B-2:1	Chr 10:43246795	A/C	0.714(0.557, 0.917)	0.729(0.542, 0.981)	0.517(0.224, 1.191)	0.705(0.530, 0.939)	0.555(0.242, 1.275)
rs77628730	lnc-CCAT1	Chr 8:128220966	A/T	1.261(1.046, 1.521)	1.206(0.936, 1.554)	1.807(1.085, 3.011)	1.273(0.997, 1.624)	1.656(1.008, 2.722)
rs6989575	lnc-CCAT1	Chr 8:128226195	C/T	1.030(0.870, 1.219)	1.200(0.902, 1.595)	1.004(0.703, 1.433)	1.141(0.871, 1.496)	0.892(0.658, 1.210)
rs7816475	lnc-CCAT1	Chr 8:128225440	A/G	1.191(0.960, 1.478)	1.435(1.097, 1.878)	0.868(0.451, 1.672)	1.358(1.049, 1.757)	0.776(0.405, 1.485)
rs6470502	lnc-CCAT1	Chr 8:128221510	T/C	0.505(0.406, 0.628)	0.329(0.244, 0.445)	0.629(0.387, 1.023)	0.382(0.292, 0.501)	0.840(0.521, 1.355)
rs1518338	lncRNA-TUSC7	Chr 3:116429325	C/G	1.084(0.890, 1.320)	1.355(1.051, 1.747)	0.635(0.347, 1.163)	1.251(0.979, 1.598)	0.561(0.309, 1.019)
rs2867837	lncRNA-TUSC7	Chr 3:116436449	G/A	0.948(0.766, 1.173)	0.697(0.524, 0.927)	1.582(0.955, 2.622)	0.827(0.637, 1.073)	1.742(1.057, 2.781)
rs12494960	lncRNA-TUSC7	Chr 3:116435140	A/C	2.616(2.122, 3.226)	2.566(1.967, 3.347)	7.672(3.790, 15.530)	2.897(2.241, 3.744)	5.392(2.681, 10.844)
rs74798803	lncRNA-CASC9	Chr 8:76136496	T/C	0.966(0.795, 1.174)	0.992(0.772, 1.274)	0.844(0.463, 1.538)	0.976(0.765, 1.245)	0.847(0.469, 1.529)
rs7818137	lncRNA-CASC9	Chr 8:76135674	T/C	1.198(1.012, 1.417)	1.581(1.169, 2.139)	1.432(0.991, 2.069)	1.539(1.152, 2.056)	1.036(0.768, 1.398)
rs550894	lncRNA-NEAT1	Chr 11:65211940	T/G	1.129(0.934, 1.364)	1.242(0.964, 1.601)	1.274(0.764, 2.124)	1.264(0.977, 1.591)	0.865(0.526, 1.423)
rs3825071	lncRNA-NEAT1	Chr 11: 65212122	A/G	1.475(1.161, 1.873)	1.687(1.278, 2.227)	1.136(0.405, 3.191)	1.654(1.260, 2.171)	0.986(0.352, 2.761)
rs580933	lncRNA-NEAT1	Chr 11:65196884	G/C	0.980(0.807, 1.191)	1.160(0.897, 1.500)	0.808(0.486, 1.343)	1.099(0.860, 1.405)	0.762(0.463, 1.253)
rs7943779	lncRNA-NEAT1	Chr 11:65194586	A/G	1.537(1.194, 1.978)	1.615(1.215, 2.147)	2.078(0.484, 8.918)	1.627(1.228, 21.56)	1.840(0.429, 7.885)
rs911157	lncRNA-NKILA	Chr 20:56286443	T/C	1.741(1.192, 2.542)	1.651(1.099, 2.480)	1.869(0.157, 22.253)	1.656(1.107, 2.477)	1.771(0.148, 21.153)
rs16981280	lncRNA-NKILA	Chr 20:56287862	C/G	0.756(0.636, 0.899)	0.677(0.519, 0.884)	0.539(0.364, 0.798)	0.646(0.500, 0.833)	0.677(0.473, 0.970)
rs2273534	lncRNA-NKILA	Chr 20:56285540	C/T	0.919(0.777, 1.087)	1.073(0.794, 1.449)	0.902(0.631, 1.291)	1.019(0.765, 1.358)	0.859(0.643, 1.148)
rs957313	lncRNA-NKILA	Chr 20:56286812	T/G	1.040(0.790, 1.370)	1.110(0.810, 1.520)	1.203(0.392, 3.687)	1.115(0.820, 1.516)	1.180(0.385, 3.609)

The unconditional logistic regression analysis adjusted by age, sex, smoking, drinking, and family history of tumors in first-degree relatives.
CI, confidence interval;; lncRNA, long noncoding RNA; OR, odds ratio; RA, risk allele; SNP, single-nucleotide polymorphism.

with the middle quintile category (40th–60th percentile) as the reference.

The OR of gastric cancer expressed in the percentile of PRS was compared with the predicted OR under the multiplicative polygene genetic model. The contribution of *Helicobacter pylori* infection, smoking, alcohol consumption, and family history to PRS was assessed individually and jointly by fitting additional interactions in the model. The empirical *P* value was used to perform 10,000 fittings within the model to optimize model parameters and build the optimal model.

Receiver operating characteristic and area under the curve (AUC) were used to evaluate the gastric cancer recognition degree of different models. The Akaike information criterion (AIC) and Bayesian information criterion (BIC) were used to evaluate the goodness of fit of risk prediction model.

RESULTS

Baseline characteristics of subjects

A total of 544 patients with gastric cancer and 544 healthy controls from the community were included in this study (Table 1). There was no significant difference in the mean age between the case group (57.80 ± 12.06) and the control group (57.02 ± 11.97) ($P = 0.284$) and no significant difference in sex and alcohol consumption between the case group and the control group ($P > 0.05$). Smoking status (56.07%), family history of tumor (11.93%), and *H. pylori* infections (63.55%) in the case group were higher than those in the control group (smoking, $P = 0.029$; family history of tumor, $P < 0.001$; *H. pylori* infection, $P = 0.014$).

Selection of lncRNA SNPs and susceptibility to gastric cancer

Based on bioinformatics method, 21 SNPs in lncRNA genes that affect the potential binding ability of microRNAs were screened (Table 2). Multivariate unconditional logistic regression analysis was used to explore the association between lncRNA SNPs and the risk of gastric cancer based on 5 genetic models (allele, heterozygosity, homozygosity, dominance, and recessive models) adjusted by sex, age, smoking status, drinking status, and family history of tumor. The results showed that 14 SNPs (rs1859168, rs4784659, rs579501, rs77628730, rs7816475, rs6470502, rs1518338, rs2867837, rs12494960, rs7818137, rs3825071, rs7943779, rs911157, and rs16981280) were significantly associated with gastric cancer risk (Table 2).

Distribution of GRS

In the distribution of GRS in cases and controls, the mean value of PRS in cases was higher than that in controls (Figure 2a). According to the PRS algorithm, the lncRNA SNPs were assigned to 1,088 subjects, and the normal test was conducted according to the frequency density distribution. The results revealed that the PRS distribution was consistent with the normal distribution (Figure 2b).

Construction of PRS risk prediction model

The bar plot was used to present that the correlation results obtained under different *P* value thresholds (P_T) correspond to

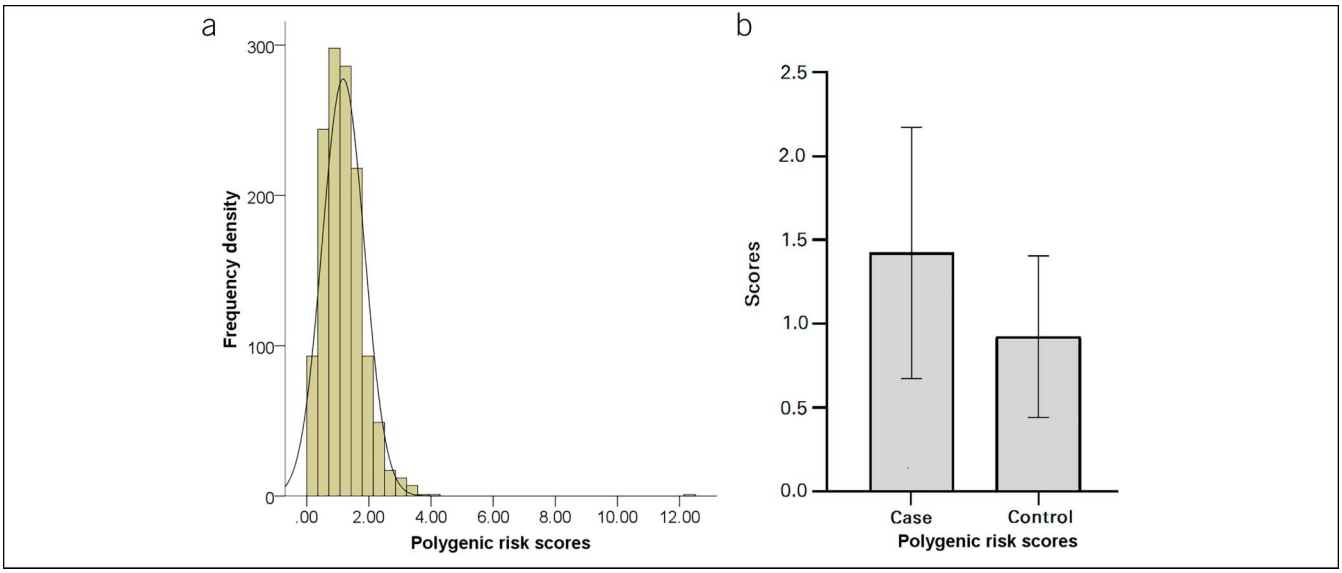

Figure 2.(**a**) Normal distribution of lncRNA SNPs PRS. (**b**) Distribution of lncRNA SNPs PRS in patients and controls. lncRNA, long noncoding RNA; PRSs, polygenic risk scores; SNPs, single-nucleotide polymorphisms.

the variance proportion of PRS interpretation, that is, the distribution of explanatory value (R^2) of estimated phenotypic variation. Figure 3a shows the R^2 value (vertical axis) of phenotypic variation of PRS model under different P_T values (horizontal axis), and the point with the highest histogram indicates that the model was optimal (when $P_T = 0.2094$); in the optimal PRS model, approximately 20.1% of the cases was caused by genetic variation ($P = 1.9 \times 10^{-34}$).

The high-resolution plot was applied to reveal the empirical P value distribution corresponding to the correlation results obtained under different P_T values. The results are shown in Figure 3b. In this model, the best P_T value was at the highest point of the broken line, and the P_T was 0.2094.

According to the distribution of the data set, the data were divided into percentile, with 40%–60% percentile as reference. Meanwhile, quantile plots were used to show the impact of PRS on phenotypic prediction risk. The results showed that the individual risk of gastric cancer decreased with the decrease of quantile. With the increase of the quantile, the risk of gastric cancer was significantly increased (Figure 4). The individual risk of gastric cancer in

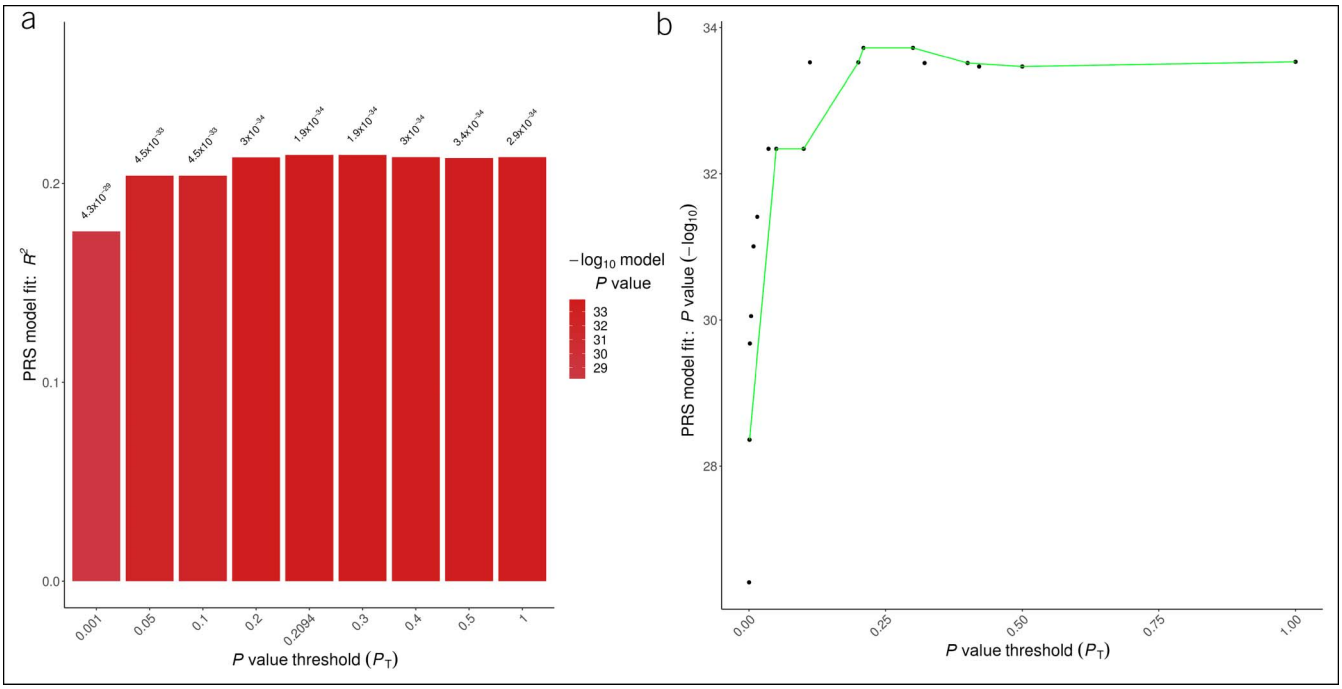

Figure 3.(**a**) PRS prediction threshold (P_T) of gastric cancer and phenotypic variation interpretation bar plot. (**b**) PRS P_T and model goodness of fit high-resolution plot. PRS, polygenic risk score.

Table 3. Regression analysis of PRS and corresponding quantile with the risk of gastric cancer

Quantile	OR	CI.U	CI.L	Group	N
0,10	0.179	0.314	0.102	1	109
10,20	0.363	0.591	0.223	2	110
20,30	0.282	0.471	0.170	3	108
30,40	0.519	0.831	0.324	4	108
40,60	1 (Ref.)	1(Ref.)	1 (Ref.)	5 (Ref.)	218
60,70	1.142	1.819	0.717	6	108
70,80	2.132	3.495	1.301	7	109
80,90	1.583	2.548	0.984	8	109
90,100	5.751	10.702	3.090	9	109

OR, odds ratio; PRS, polygenic risk score.

the lowest 10% of PRS was 82.1% (95% confidence interval [CI]: 0.102, 0.314) lower than that of the general population. The risk of gastric cancer in the highest 10% of PRS was 5.75 folds that of the general population (95% CI: 3.09, 10.70) (Table 3).

Evaluation of risk prediction model

According to the receiver operating characteristic curve, AUC results showed that the introduction of family history of tumor and *H. pylori* infection on the basis of PRS could significantly improve the recognition ability of the model (Table 4). By introducing different factors to compare AIC and BIC, on the basis of PRS, the introduction of family history of cancer, *H. pylori* infection, and smoking, drinking model was better than genetic risk model. Among them, the model of PRS introducing family history of tumor had the best fit (AIC = 78.14, BIC = 78.04) (Table 4).

DISCUSSION

The carcinogenic mechanism of gastric cancer has not been fully elucidated. Approximately 98% of the human genome is composed of non-coding DNA (9), and about 70% of the genome is actively transcribed, and 2% encodes known protein coding genes (10). Studies have deidentified a 9-lncRNA signature could act as a potential prognostic biomarker in the prediction of gastric cancer (11), and 18 lncRNAs were significantly associated with the survival of gastric cancer (12).

Cancer has complex molecular characteristics. Therefore, a single lncRNA expression pattern may not be sufficient to accurately predict the prognosis of gastric cancer. Nevertheless, previous studies have shown that the combination of multiple potential lncRNA biomarkers can improve the accuracy of prediction (13,14).

So far, PRS has been used to construct risk prediction models for many complex diseases. The results of a study by Rudolph et al. (15) combined PRS and environmental factors and showed that breast cancer associated SNPs interacted with environmental risk factors, and the model improved the ability of risk prediction. In other breast cancer related studies, the method of using PRS to construct risk prediction model has also obtained the similar conclusions (16–18). Meanwhile, this method has also been used in the construction of risk prediction models of psoriasis (19), stroke (20), bipolar disorder and mental disorder (21), and achieved good prediction results.

As far as we know, there is no report on the construction of risk model for gastric cancer by using PRS. The models based on GRSs are based on the genetic sites screened by GWAS or evidence-based medicine (22–24). In the study of risk prediction model, there was no study on lncRNA related SNPs, but studies have confirmed that lncRNA-related SNPs were related to gastric cancer (25–27).

In this study, bioinformatics methods were used to screen gastric cancer-related lncRNAs and verify them in the population after functional prediction. Based on correlation verification results, a panel of 21 lncRNA SNPs combined with data on classic risk factors further stratified individual gastric cancer risk in the population. After adjusting for classical factors, the relative risk of polygenic score was well calibrated.

The association between 21 candidate functional SNPs and susceptibility to gastric cancer was validated in Chinese population. Multivariate unconditional logistic regression analysis showed that 14 lncRNA SNPs were statistically related to the risk of gastric cancer. Among them, some associated SNPs have been confirmed in our previous studies (28).

The associated lncRNA SNPs were put into the prediction models and empirical *P*-value was used to perform 10,000 fittings within the model to optimize model parameters and construct the optimal model. About 20.1% of the cases were caused by genetic variation in the optimal PRS model. This indicator estimated the proportion of variance explained by PRS, which assumed that the underlying variable was normally distributed (29,30). With the 40%–60% percentile as a reference, the results showed that with the decrease of the quantile, the individual risk of gastric cancer showed a downward trend, and the risk of gastric cancer in the highest 10% of PRS was 5.75-fold that of the general population (95% CI: 3.09, 10.70).

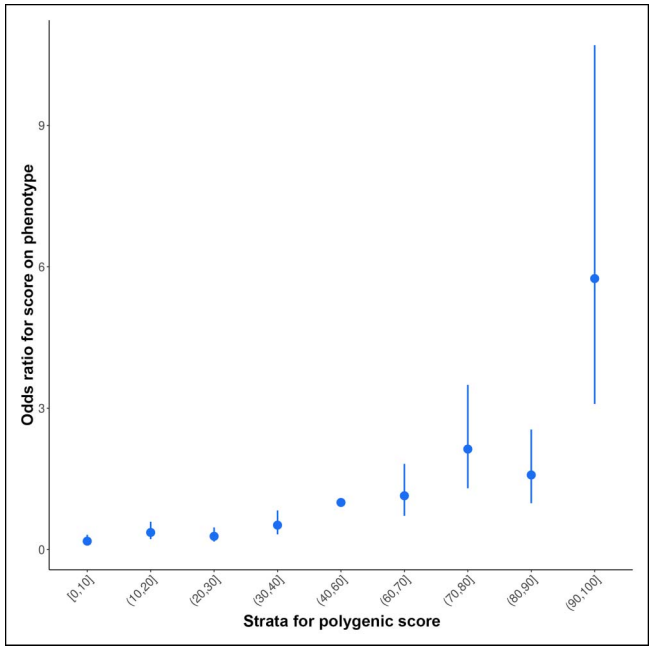

Figure 4. Quantile diagram of polygenic risk score phenotype prediction risk.

Table 4. Comparison of goodness of fit model and risk identification

Model and variables	AUC (95% CI)	AIC	BIC
PRS	0.737 (0.71, 0.76)	1727.91	1738.28
PRS + *Helicobacter pylori* infection	0.752 (0.690, 0.814)	82.08	86.55
PRS + TFH	0.773 (0.702, 0.843)	78.14	78.04
PRS + drinking	0.723 (0.67, 0.78)	460.98	468.72
PRS + smoking	0.681 (0.63, 0.73)	629.12	637.29
PRS + smoking + drinking	0.704 (0.64, 0.77)	335.43	342.45
PRS + Drinking + *H. pylori* infection	0.722 (0.635, 0.808)	174.94	180.79
PRS + Smoking + *H. pylori* infection	0.738 (0.650, 0.826)	161.01	166.71
PRS + Smoking + Drinking + *H. pylori* infection	0.735 (0.63, 0.84)	121.25	126.33

AIC, Akaike information criterion; AUC, area under the curve; BIC, Bayesian information criterion; PRS, polygenic risk score; TFH, tumor family history.

This is consistent with the results and trends of other diseases based on PRS in Chinese population (31,32). The result of recognition ability analysis showed that the introduction of family history of tumor (AUC, 95% CI: 0.752, 0.69–0.814) and *H. pylori* infection (AUC, 95% CI: 0.773, 0.702–0.843) on the basis of PRS could significantly improve the recognition ability of the model. The discrimination ability of the model was higher than that of the risk prediction model constructed by common SNP, even without the introduction of the abovementioned 2 factors (7,33,34). The PRS introducing family history of tumor had the best fit (AIC = 78.14, BIC = 78.04). The cumulative effect of this PRS with environmental and/or biological factors has been confirmed in other cancers (31,35,36).

This study has some limitations. First, the interaction among lncRNA SNPs was not dealt with in genetic association analysis, which may have a certain impact on the construction of the model. Second, this study differs from the previous analysis strategy of identifying disease-related SNPs based on GWAS. This study integrates a variety of databases, reannotates and remines high-throughput microarray, and is based on bioinformatics and Chinese population data. Moreover, there were differences in genetic background between verification population and chip data population, which may lead to a certain degree of bias in the validation results. Third, the molecular mechanism and function of these lncRNAs are still unclear and need to be further studied.

In summary, our results demonstrated the potential value of lncRNA SNPs in the prevention of gastric cancer risk, especially in identifying individuals at higher risk of gastric cancer. Therefore, it can be used as an accurate and cost-effective initial large-scale prescreening tool to improve the level of primary prevention of gastric cancer. A large-scale population screening program should be launched to test its feasibility in the future.

CONFLICTS OF INTEREST

Guarantors of the article: Fu-Jiao Duan, PhD and Kai-Juan Wang, PhD.

Specific author contributions: F.-J.D. and K.-J.W.: study concept and design. H.Y. and C.-H.S.: analysis of data. P.W.: interpretation of data. L.-P.D. and J.-Y.Z: drafting the manuscript. All authors reviewed this manuscript and approved the final draft.

Financial support: This work was funded by the National Natural Science Foundation of China (81373097), the National Science and Technology Major Project of China (2018ZX10302205), and Joint construction project of medical Science and Technology in Henan Province (LHGJ20210184).

Potential competing interests: None to report.

Study Highlights

WHAT IS KNOWN

- ✓ Polygenic risk scores (PRSs) have been developed and are increasingly used for different cancer risk stratification.
- ✓ Studies have confirmed that long noncoding RNA (lncRNA) single nucleotide polymorphisms (SNPs) are associated with the risk of gastric cancer.
- ✓ The lncRNA SNPs are not involved in the current construction of cancer-related risk models.

WHAT IS NEW HERE

- ✓ In the distribution of genetic risk score in cases and controls, the mean value of PRS in cases was higher than that in controls.
- ✓ Approximately 20.1% of the cases was caused by genetic variation in optimal PRS model.
- ✓ The introduction of family history of tumor and *Helicobacter pylori* infection on the basis of PRS could significantly improve the recognition ability of the model.

REFERENCES

1. Torre LA, Siegel RL, Ward EM, et al. Global cancer incidence and mortality rates and trends–An update. Cancer Epidemiol Biomarkers Prev 2016;25:16–27.
2. Ferlay J, Soerjomataram I, Dikshit R, et al. Cancer incidence and mortality worldwide: Sources, methods and major patterns in GLOBOCAN 2012. Int J Cancer 2015;136:E359–86.
3. Kitayama J, Ishigami H, Yamaguchi H, et al. Treatment of patients with peritoneal metastases from gastric cancer. Ann Gastroenterol Surg 2018;2:116–23.
4. Zhao J, Liu Y, Zhang W, et al. Long non-coding RNA Linc00152 is involved in cell cycle arrest, apoptosis, epithelial to mesenchymal transition, cell migration and invasion in gastric cancer. Cell Cycle 2015; 14:3112–23.

5. Rinn JL, Chang HY. Genome regulation by long noncoding RNAs. Annu Rev Biochem 2012;81:145–66.
6. Mattick JS, Makunin IV. Non-coding RNA. Hum Mol Genet 2006;15(1): R17–29.
7. Hachiya T, Kamatani Y, Takahashi A, et al. Genetic predisposition to ischemic stroke: A polygenic risk score. Stroke 2017;48:253–8.
8. Chatterjee N, Shi J, García-Closas M. Developing and evaluating polygenic risk prediction models for stratified disease prevention. Nat Rev Genet 2016;17:392–406.
9. Mattick JS. Non-coding RNAs: The architects of eukaryotic complexity. EMBO Rep 2001;2:986–91.
10. Djebali S, Davis CA, Merkel A, et al. Landscape of transcription in human cells. Nature 2012;489:101–8.
11. Cai C, Yang L, Tang Y, et al. Prediction of overall survival in gastric cancer using a nine-lncRNA. DNA Cell Biol 2019;38:1005–12.
12. Gao S, Zhao ZY, Wu R, et al. Prognostic value of long noncoding RNAs in gastric cancer: A meta-analysis. Onco Targets Ther 2018;11:4877–91.
13. Ke D, Li H, Zhang Y, et al. The combination of circulating long noncoding RNAs AK001058, INHBA-AS1, MIR4435-2HG, and CEBPA-AS1 fragments in plasma serve as diagnostic markers for gastric cancer. Oncotarget 2017;8:21516–25.
14. Wu B, Wang K, Fei J, et al. Novel threelncRNA signature predicts survival in patients with pancreatic cancer. Oncol Rep 2018;40:3427–37.
15. Rudolph A, Song M, Brook MN, et al. Joint associations of a polygenic risk score and environmental risk factors for breast cancer in the breast cancer association consortium. Int J Epidemiol 2018;47:526–36.
16. Maas P, Barrdahl M, Joshi AD, et al. Breast cancer risk from modifiable and nonmodifiable risk factors among white women in the United States. JAMA Oncol 2016;2:1295–302.
17. Shieh Y, Hu D, Ma L, et al. Breast cancer risk prediction using a clinical risk model and polygenic risk score. Breast Cancer Res Treat 2016;159: 513–25.
18. Lee A, Mavaddat N, Wilcox AN, et al. BOADICEA: A comprehensive breast cancer risk prediction model incorporating genetic and nongenetic risk factors. Genet Med 2019;21:1708–18.
19. Yin X, Cheng H, Lin Y, et al. A weighted polygenic risk score using 14 known susceptibility variants to estimate risk and age onset of psoriasis in Han Chinese. PLoS One 2015;10:e0125369.
20. Reginsson GW, Ingason A, Euesden J, et al. Polygenic risk scores for schizophrenia and bipolar disorder associate with addiction. Addict Biol 2018;23:485–92.
21. Power RA, Steinberg S, Bjornsdottir G, et al. Polygenic risk scores for schizophrenia and bipolar disorder predict creativity. Nat Neurosci 2015; 18:953–5.
22. Li J, Miu X, Lin D. Genome wide association of common tumors in China. Chinese J Nat 2015;37:1–7.
23. Yan C, Zhu M, Ding Y, et al. Meta-analysis of genome-wide association studies and functional assays decipher susceptibility genes for gastric cancer in Chinese populations. Gut 2020;69:641–51.
24. Tian J, Miu X, Lin D. Research Progress on genetic risk prediction models of common malignant tumors in Chinese population. J Biotechnol 2016;6:10–15.
25. Zhu M, Wang Y, Liu X, et al. LncRNAs act as prognostic biomarkers in gastric cancer: A systematic review and meta-analysis. Front Lab Med 2017;1:59–68.
26. Fattahi S, Kosari-Monfared M, Golpour M, et al. LncRNAs as potential diagnostic and prognostic biomarkers in gastric cancer: A novel approach to personalized medicine. J Cel Physiol 2020;235:3189–206.
27. Esfandi F, Salehnezhad T, Taheri M, et al. Expression assessment of a panel of long non-coding RNAs in gastric malignancy. Exp Mol Pathol 2020;113:104383.
28. Duan F, Jiang J, Song C, et al. Functional long non-coding RNAs associated with gastric cancer susceptibility and evaluation of the epidemiological efficacy in a central Chinese population. Gene 2018;646:227–33.
29. Lee SH, Goddard ME, Wray NR, et al. A better coefficient of determination for genetic profile analysis. Genet Epidemiol 2012;36:214–24.
30. Lee SH, Wray NR, Goddard ME, et al. Estimating missing heritability for disease from genome-wide association studies. Am J Hum Genet 2011;88:294–305.
31. Dai J, Lv J, Zhu M, et al. Identification of risk loci and a polygenic risk score for lung cancer: A large-scale prospective cohort study in Chinese populations. Lancet Respir Med 2019;7:881–91.
32. Khera AV, Chaffin M, Aragam KG, et al. Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. Nat Genet 2018;50:1219–24.
33. Dudbridge F, Pashayan N, Yang J. Predictive accuracy of combined genetic and environmental risk scores. Genet Epidemiol 2018;42:4–19.
34. Hsieh YC, Tu SH, Su CT, et al. A polygenic risk score for breast cancer risk in a Taiwanese population. Breast Cancer Res Treat 2017;163:131–8.
35. Lakeman IMM, Hilbers FS, Rodríguez-Girondo M, et al. Addition of a 161-SNP polygenic risk score to family history-based risk prediction: Impact on clinical management in non-BRCA1/2 breast cancer families. J Med Genet 2019;56:581–9.
36. Vachon CM, Scott CG, Tamimi RM, et al. Joint association of mammographic density adjusted for age and body mass index and polygenic risk score with breast cancer risk. Breast Cancer Res 2019;21:68.