Bayesian Survival Analysis of Genetic Variants in PTPRN2 Gene for Age at Onset of Cancer

Ke-Sheng Wang  
*East Tennessee State University*

Yue Pan

Weize Wang

Chun Xu  
*The University of Texas Rio Grande Valley*

Follow this and additional works at: [https://scholarworks.utrgv.edu/hbs_fac](https://scholarworks.utrgv.edu/hbs_fac)

Part of the Medicine and Health Sciences Commons

**Recommended Citation**


This Article is brought to you for free and open access by the College of Health Professions at ScholarWorks @ UTRGV. It has been accepted for inclusion in Health & Biomedical Sciences Faculty Publications and Presentations by an authorized administrator of ScholarWorks @ UTRGV. For more information, please contact justin.white@utrgv.edu, william.flores01@utrgv.edu.
Bayesian Survival Analysis of Genetic Variants in PTPRN2 Gene for Age at Onset of Cancer

Ke-Sheng Wang1,*, Yue Pan2, Weize Wang3 and Chun Xu4

1Department of Biostatistics and Epidemiology, East Tennessee State University, USA
2Department of Public Health Sciences, University of Miami, USA
3Department of Biostatistics, Florida International University, USA
4Department of Pediatrics, Texas Tech University Health Sciences Center, USA

*Corresponding author: Kesheng Wang, Department of Biostatistics and Epidemiology, College of Public Health, East Tennessee State University, PO Box 70259, Lamb Hall, Johnson City, TN 37614-1700, USA, Tel:1-423-439-4401, Fax: 1-423-439-4606, E-mail: wangk@etsu.edu

Abstract

Background: The protein tyrosine phosphatase, receptor type, N polypeptide 2 (PTPRN2) gene may play a role in cancer; however, no study has focused on the associations of genetic variants within the PTPRN2 gene with age at onset (AAO) of cancer.

Methods: This study examined 220 single nucleotide polymorphisms (SNPs) within the PTPRN2 gene in the Marshfield sample with 716 cancer cases (any diagnosed cancer, excluding minor skin cancer) and 2,846 non-cancer controls. Multiple logistic regression model and linear regression model in PLINK software were used to examine the association of each SNP with the risk of cancer and AAO, respectively. For survival analysis of AAO, both classic Cox regression and Bayesian survival analysis using the Cox proportional hazards model in SAS v. 9.4 were applied to detect the association of each SNP with AAO. The hazards ratios (HRs) with 95% confidence intervals (CIs) were estimated.

Results: Single marker analysis identified 10 SNPs associated with the risk of cancer and 9 SNPs associated with AAO (p < 0.05). SNP rs7783909 revealed the strongest association with cancer (p = 6.52x10^{-3}); while the best signal for AAO was rs4909140 (p = 6.18x10^{-4}), which was also associated with risk of cancer (p = 0.0157). Classic Cox regression showed that 11 SNPs were associated with AAO (top SNP rs4909140 with HR = 1.11-1.71, p = 3.3x10^{-4}). Bayesian Cox regression model showed similar results to those using the classic Cox regression (top SNP rs4909140 with HR = 1.39, 95%CI = 1.1-1.69).

Conclusions: This study provides evidence of several genetic variants within the PTPRN2 gene influencing the risk of cancer and AAO, and will serve as a resource for replication in other populations.

Keywords

Cancer, Age at onset, PTPRN2, Single nucleotide polymorphism, Bayesian analysis, Survival analysis

Introduction

Cancer continues to remain a significant public health issue globally. It is the leading cause of death in both developed and emerging economies [1]. In 2012, there were 14.1 million new cancer cases, 8.2 million cancer mortalities, and 32.6 million people living with cancer globally [2]. Cancers are caused by a complex interplay between genetic predisposition and environment. Family and twin studies have shown the estimated effects of heritability of the colorectal cancer (35%) [3], breast cancer (25-30%) [3-5], prostate cancer (42-58%) [3,6], and lung cancer (25-26%) [3,7].

The protein tyrosine phosphatase, receptor type, N polypeptide 2 (PTPRN2) gene (also known as IAR, ICAAR, PTPRP, IA-2-beta, R-PTP-N2) is located at 7q36 [8,9]. The PTPRN2 is expressed predominantly in human brain and pancreas and in mouse brain, pancreas, and insulinoma cell lines [10,11]. One methylation study found that PTPRN2 showed highly significant hypermethylation in squamous cell lung cancer tissue [12]. Another gene expression profile study suggested that PTPRN2 was associated with metastatic prostate cancer [13]. A recent study reported that PTPRN2 was expressed predominantly in endocrine and neuronal cells, where it might function in exocytosis and suggested as a novel candidate biomarker and therapeutic target in breast cancer [14].

However, no study has focused on the associations of genetic variants of PTPRN2 gene with age at onset (AAO) of cancer. Furthermore, Bayesian methods have become increasingly popular in many areas of scientific research including genetic association studies, which may have some advantages in flexibility, and incorporating information from previous studies and dealing with sparse-data [15-17]. In this study, we explored the associations of 220 SNPs in the PTPRN2 gene with AAO of cancer using a Bayesian survival analysis in a Caucasian sample.

Subjects and Methods

The Marshfield sample

The Marshfield sample is from the publicly available data in A Genome-Wide Association Study on Cataract and HDL in the...
Personalized Medicine Research Project Cohort - Study Accession: phs000170.p1 (dbGaP). The details about these subjects were described elsewhere [18,19]. Cancer cases were defined as any diagnosed cancer excluding minor skin cancer; while AAO cancer was defined by date of the earliest cancer diagnosis in the registry. Social factors used in this study were age, gender, alcohol use in the past month (yes or no), and smoking status (never smoking, current smoking and past smoking). Obesity was determined as a body mass index (BMI) ≥ 30. Genotyping data using the ILLUMINA Human660W-Quad_v1_A are available for 3562 Caucasian individuals (716 cancer cases and 2848 controls). Within the PTPRN2 gene, 220 SNPs were available and therefore included in the analysis.

Statistical methods

Linear and logistic regression models in PLINK software: The categorical variables were presented as frequencies and percentages. The continuous variables were reported as the means ± standard deviation. Quality-control and association analyses were implemented using PLINK V1.07 [20]. First, Hardy-Weinberg equilibrium (HWE) was tested for all the SNPs using the controls; then, minor allele frequency (MAF) was determined for each SNP. Multiple logistic regression analysis of each SNP with risk of cancer as a binary trait, adjusted for sex, age*age, alcohol use, smoking status, and obesity, was performed using PLINK; while the asymptotic p-values were observed and the odds ratio (OR) and 95% confident interval (CI) were estimated. Furthermore, AAO was firstly log transformed, then multiple linear regression analysis of each SNP with log transformed AAO of cancer, adjust for sex, alcohol use, smoking status, and obesity was performed; while the asymptotic p-values were observed and the regression coefficient (β) and 95% CI were estimated. To control for type I errors arising from multiple hypothesis testing, a false discovery rate (FDR) was defined in Benjamini and Hochberg [21] as the expected proportion of false discoveries. In addition, empirical p-values were generated by 100,000 permutation tests using Max (T) permutation procedure in this procedure, pointwise estimate of an individual SNP’s significance (empirical pointwise p-values) was calculated.

Cox proportional hazards models in PROC PHREG: The proportional hazards model or Cox regression model, is widely used in the analysis of time-to-event data to explain the effect of explanatory variables on hazard rates. The PHREG procedure fits the Cox model by maximizing the partial likelihood function; this eliminates the unknown baseline hazard and accounts for censored survival times. In the Bayesian approach, the partial likelihood function is used as the likelihood function in the posterior distribution [22]. In the non-Bayesian analysis, the Akaake information criterion (AIC) was used as a measure of goodness of model fit that balances model fit against model simplicity [23,24]. Bayesian Cox regression can be requested by using the BAYES statement in the PHREG procedure. A Markov chain Monte Carlo (MCMC) method by Gibbs sampling was used to simulate samples from the posterior distribution. In a Bayesian analysis, a Gibbs chain of samples from the posterior distribution was generated for the model parameters. Summary statistics (mean, standard deviation, quartiles, the highest posterior density (HPD) and credit disparity correlation matrix) and convergence diagnosables (Geweke; the effective sample size; and Monte Carlo standard errors) were computed for each parameter, as well as the correlation matrix and the covariance matrix of the posterior sample. Trace plots, posterior density plots, and autocorrelation function plots were created for each parameter [17]. The hazards ratios (HRs) with 95% CIs were estimated.

For the present study of the AAO, the normal prior was chosen for the coefficients. In Bayesian analysis, a deviance information criterion (DIC) is available for model comparison instead of AIC. DIC is a hierarchical modeling generalization of the AIC; while DIC is intended as a generalization of AIC [25]. The following program showed one SNP rs4909140, sex, alcohol use, smoking status, and obesity with the AAO of cancer. The rs4909140 has 3 genotypes - G_G, G_T and T_T, respectively; while the T_T genotype was considered as the reference.

Table 1: Descriptive characteristics of cases and controls

<table>
<thead>
<tr>
<th></th>
<th>Non-Cancer</th>
<th>Cancer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number</td>
<td>2848</td>
<td>716</td>
</tr>
<tr>
<td>Sex, N (%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Males</td>
<td>1135(40%)</td>
<td>340(47%)</td>
</tr>
<tr>
<td>Females</td>
<td>1713(60%)</td>
<td>376(53%)</td>
</tr>
<tr>
<td>Obesity, N (%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>No</td>
<td>1700(80%)</td>
<td>422(59%)</td>
</tr>
<tr>
<td>Yes</td>
<td>1148(40%)</td>
<td>294(41%)</td>
</tr>
<tr>
<td>Alcohol, N (%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>No</td>
<td>1060(37%)</td>
<td>288(40%)</td>
</tr>
<tr>
<td>Yes</td>
<td>1783(63%)</td>
<td>425(60%)</td>
</tr>
<tr>
<td>Smoking, N (%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Never</td>
<td>1487(52%)</td>
<td>327(46%)</td>
</tr>
<tr>
<td>Current</td>
<td>254(9%)</td>
<td>54(7%)</td>
</tr>
<tr>
<td>Past</td>
<td>1104(39%)</td>
<td>331(47%)</td>
</tr>
<tr>
<td>Age, years</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean ± SD</td>
<td>65.1 ± 11.3</td>
<td>71.1 ± 10.3</td>
</tr>
<tr>
<td>Range</td>
<td>46-90</td>
<td>46-90</td>
</tr>
<tr>
<td>Age at onset, years</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean ± SD</td>
<td>-</td>
<td>64.2 ± 12.8</td>
</tr>
<tr>
<td>Range</td>
<td>-</td>
<td>23-90</td>
</tr>
</tbody>
</table>

Descriptive statistics and Cox regression analyses were conducted with SAS V.9.4 (SAS Institute, Cary, NC, USA).

Results

Genotype quality control and descriptive statistics

We removed 1 SNP with HWE p < 10^ -4. All other 219 SNPs were in HWE with MAF > 1% in the controls. The demographic characteristics of the subjects in the study are presented in Table 1. There were slightly more females than males in both cases and controls. The age ranged from 46 to 90 years and AAO of cancer ranged from 23 to 90 years.

Multiple linear and logistic regression analyses using PLINK

Using a single marker analysis, we identified 10 SNPs associated with the risk of cancer and 9 SNPs associated with AAO (p < 0.05) in the Marshfield sample (Table 2). SNP rs7783909 revealed the strongest association with cancer (p = 6.52x10^-3); while the best signal for AAO was rs4909140 (p = 6.18x10^-4), which was also associated with risk of cancer (p = 0.0157). For the 10 SNPs associated with risk of cancer, the FDR was 90%; while the FDR for the two AAO mostly associated SNPs (rs4409140 and rs1670340) were 21% and 39%, respectively. Furthermore, we conducted a permutation test in PLINK and found that all the cancer and or AAO associated SNPs had empirical point wise p-values p < 0.05 using a permutation test (Table 2).

Classic and Bayesian Cox regression analyses using PROC PHREG

Classic Cox regression model showed that 11 SNPs were associated with AAO (top SNP rs4909140 with HR=1.38, 95%CI = 1.11-1.71, p = 3.3x10^-3). The HRs based on the Bayesian survival analyses revealed similar results to those using the non-Bayesian analyses results (Table 3). The DIC for the 11 SNPs in the Bayesian analyses were similar to
The trace plot, posterior density plot, and autocorrelation function plot based on Bayesian analysis (Figure 1) indicated that the Markov chain had stabilized with good mixing for rs4909140. The posterior density plot, which estimates the posterior marginal distributions for the 7 regression coefficients showed a smooth, unimodal shape for the posterior marginal distribution (Figure 2). Table 4 shows the posterior summary of rs4909140 with HR = 1.39, 95% CI = 1.1-1.69.

**Discussion**

In this study, we identified 10 SNPs associated with the risk of cancer and 9 SNPs with AAO using the PLINK software and 11 SNPs...
revealed associations with AAO using Cox survival model in SAS. Bayesian Cox regression model revealed similar findings to those using the classic Cox regression. To our knowledge, this is the first candidate gene study to provide evidence of several genetic variants within the PTPRN2 gene associated with the risk of cancer and AAO.

Previous studies have showed that PTPRN2 is an autoantigen for type 1 diabetes which is an insulin-dependent diabetes mellitus and autoimmune disease; while PTPRN2 is reactive with type 1 diabetes patient sera and is likely to be an islet cell antigen useful in the preclinical screening of individuals for the risk of type 1 diabetes.
Conclusion

This study provides evidence of several genetic variants within the PTPRN2 gene influencing the risk and AAO of cancer. Future functional study of this gene may help to better characterize the genetic architecture of cancers.

Role of the funding sources

No founding source is given for the present paper.

Disclosure

All authors have reported no financial interests or potential conflicts of interest.

Acknowledgement

Funding support for the Personalized Medicine Research Project (PMRP) was provided through a cooperative agreement (U01HG004608) with the National Human Genome Research Institute (NHGRI), with additional funding from the National Institute for General Medical Sciences (NIGMS). The samples used for PMRP analyses were obtained with funding from Marshfield Clinic, Health Resources Service Administration Office of Rural Health Policy grant number D1A RH0005, and Wisconsin Department of Commerce Technology Development Fund contract number TDF FY010718. Funding support for genotyping, which was performed at Johns Hopkins University, was provided by the NIH (U1HG004438). Assistance with phenotype harmonization and genotype cleaning was provided by the EMERGE Administrative Coordinating Center (U01HG004603) and the National Center for Biotechnology Information (NCBI). The datasets used for the analyses described in this manuscript were obtained from dbGaP at http://www.ncbi.nlm.nih.gov/gap through dbGaP accession number phs000170.v1.p1. This study was approved by the Internal Review Board (IRB), East Tennessee State University.

References


