SciDoc Publishers
Infer, Interpret & Inspire Science

International Journal of Bioinformatics & Biological Systems (IJBBS)

Model Selection in Survival Analysis of EXT2 Gene Polymorphisms with Age at Onset of Type 2 Diabetes

Research Article

Wang KS[1*], Liu Y[1], Gong S[1], Pan Y[2], Wang L[1], Xie C[3]

[1] Department of Biostatistics and Epidemiology, College of Public Health, East Tennessee State University, Johnson City, TN 37614, USA.
[2] Department of Public Health Sciences, Miller School of Medicine, University of Miami, Miami, FL 33124, USA.
[3] Division of Biostatistics and Bioinformatics, Department of Environmental Health, University of Cincinnati, Cincinnati, OH 45267, USA.

Abstract

Background: This study aimed to compare the Cox regression and parametric survival models in genetic association analysis of age at onset (AAO) of type 2 diabetes (T2D) and examine the effect of exostosin 2 (EXT2) gene on risk and AAO of T2D.

Methods: We tested 22 single nucleotide polymorphisms (SNPs) within the EXT2 gene in the Marshfield sample among 878 T2D cases and 2,686 non-diabetes controls. Multiple logistic and linear regression models in PLINK software were used to examine the association of each SNP with the risk of T2D and AAO of T2D, respectively. Cox regression in PROC PHREG and parametric survival models (including exponential, Weibull, log-normal, log-logistic and gamma models) in PROC LIFEREG in SAS 9.4 were used to perform survival analysis of AAO. The Akaike information criterion (AIC) and Bayesian information criterion (BIC) were used to compare the competing models.

Results: PLINK software initially identified 1 SNP associated with the risk of T2D (rs7111879 with $p=6.33 \times 10^{-3}$) and 3 SNPs associated with AAO (rs7111879, rs42376464 and rs4755230 with $p=3.26 \times 10^{-2}$, 5.79×10^{-5} and 2.76×10^{-5}, respectively). AIC values showed that the gamma distribution is the best model for above 3 SNPs and followed by the Weibull distribution; whereas BIC criteria showed that the gamma distribution is similar to the Weibull distribution.

Conclusion: This study reveals that the parametric gamma and Weibull models performed better than Cox regression in genetic association of AAO of T2D and provides evidence of several genetic variants within the EXT2 gene associated with the risk and AAO of T2D.

Keywords: Type 2 Diabetes; Age at onset; EXT2; Survival Analysis; Model Selection; Cox Regression; Parametric Models.

Introduction

Globally, there were 284.6 million of patients with diabetes in 2010 and it was predicted to be 438.4 million in 2025, including 90-95% type 2 diabetes (T2D) [1]. In the Unites States (US), the estimated prevalence of T2D is 8.3 % in adults (about 25.8 million) [2]. Individuals with T2D have higher risk for cardiovascular disease and complications [3]; while T2D is also associated with and/or has comorbidity with multiple cancers such as endometrial and prostate cancers [4-8]. T2D is a complex trait caused by a complex interplay between genetic and the environment factors. Previous twin study provided evidence that genetic factors contribute to the development of T2D [9] while the heritability of T2D is about 31–69% [10].

Cox model and Weibull proportional hazard regression model have been used to analyze incident diabetes [11-17]. To date, few studies have focused on survival analyses of genetic variants with age at onset (AAO) of T2D. One previous study examined the associations of genetic variants within alpha2B adrenoceptor gene with AAO of T2D using a multiple linear regression [18]. In another study, the Cox model was used to check the associations of transcription factor 7-like 2 (TCF7L2) gene and its upstream region with AAO of T2D in Mexican Americans [19]. In addition, the Mann-Whitney and the Kruskall-Wallis tests were used to test the associations of HNF1A gene with AAO of T2D in

***Corresponding Author**:
Dr. Ke-Sheng 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 4481
Fax : +1 423 439 4606
E-mail: wangk@etsu.edu

Received: February 9, 2016
Accepted: March 25, 2016
Published: April 13, 2016

Citation: Wang KS et al. (2016) Model Selection in Survival Analysis of EXT2 Gene Polymorphisms with Age at onset of Type 2 Diabetes. *Int J Bioinform Biol Syst.* 1(1), 1-9.

patients with maturity-onset diabetes of the young (MODY)-3 [20]. However, to the best of our knowledge, no study has used parametric survival models (including exponential, Weibull, log-normal, log-logistic, and gamma models) in genetic association analysis of AAO of T2D or compared Cox regression and parametric survival models.

The exostosin glycosyltransferase 2 (EXT2) gene (also known as SOTV; SSMS) is located at 11p11.2 [21-25]. A genome-wide association study (GWAS) in a French case-control cohort identified several novel loci for the risk of T2D including 3 single nucleotide polymorphisms (SNPs) (rs3740878, rs1113132 and rs11037909) within the EXT2 gene [26]. However, the associations were not replicated in the following cohorts including: Japanese [27], an African American case-control sample [28], the Diabetes Prevention Program (DPP) data [29], Northern European populations [30], Chinese [31-32], Mexican [33], or Lebanese Arabs [34]. Recently, a meta-analysis revealed that these three SNPs in the EXT2 were significantly associated with the risk of T2D [35]. Furthermore, another meta-analysis of 20 studies in Han Chinese confirmed the association of EXT2 gene with T2D [36]. However, no study has examined the effect of EXT2 gene on AAO of T2D.

The aim of this study is twofold: (1) to examine the associations of EXT2 gene polymorphisms with AAO in a Caucasian sample and (2) to find the best model by comparing the semi-parametric Cox regression and parametric survival models in survival analysis of AAO of T2D.

Materials and Methods

Study subjects

The Marshfield sample is from the publicly available data from A Genome-Wide Association Study on Cataract and HDL in the Personalized Medicine Research Project Cohort-Study Accession: phs000170.v1.p1 (dbGaP). The primary goals of this project are to develop and validate electronic phenotyping algorithms, to accurately identify cases and controls while maintaining a positive predictive value (PPV) of > 95%, and to conduct a genome-wide association study that advances the understanding of two specific yet interrelated disease states, while simultaneously engaging the community in these research efforts. The details about these subjects were described elsewhere [37-38]. Social and behavioral factors used in this study were age, gender, alcohol use in the past month (yes or no), BMI, and smoking status (never smoking, current smoking or past smoking). Genotyping data using the ILLUMINA Human660W-Quad_v1_A are available. The genotypes of 22 SNPs within the EXT2 gene were available in this data.

Statistical methods

Descriptive statistics and genotype quality control

Descriptive statistics were used to characterize participants' sex, BMI, alcohol use, smoking, age and AAO of T2D stratified by T2D case and control status. Hardy-Weinberg equilibrium (HWE) was tested for all 22 SNPs using the controls; then, minor allele frequency (MAF) was determined for each SNP by using PLINK v1.07 [39]. To deal with population stratification, the principal-component analysis approach [40] in HelixTree software was used to identify and exclude outlier individuals [41]. Consequently, 3564 Caucasian individuals were included (878 individuals with T2D and 2686 non-T2D individuals).

Multiple logistic and linear regression models in PLINK software

Multiple logistic regression analysis (1) of each SNP with the risk of T2D as a binary outcome, adjusted for sex, age, alcohol use, smoking status and BMI, was performed using PLINK; while the asymptotic p-values were observed and the odds ratio (OR) and 95% confident interval (CI) were estimated.

$$\text{logit}(p(Y_1=1)) = \beta_0 + \beta_1 SNP_k + \beta_2 Sex + \beta_3 Age + \beta_4 Alcohol + \beta_5 Smoking + \beta_6 BMI \quad (1)$$

where Y_1 is T2D (Y_1=1 if T2D) and SNP_k is the genotype at the k^{th} SNP.

The similar procedure was performed for the multiple linear regression analysis (2) of each SNP with the AAO of T2D as a continuous outcome.

$$Y = \beta_0 + \beta_1 SNP_k + \beta_2 Sex + \beta_3 Alcohol + \beta_4 Smoking + \beta_5 BMI \quad (2)$$

where Y is AAO of T2D and SNP_k is the genotype at the k^{th} SNP.

Multiple testing

Bonferroni correction ($\alpha=0.05/22=2.27\times10^{-3}$) was used for statistical significance [42]. In addition to obtain nominal Type I error rate, empirical *p*-values were generated by 100,000 permutation tests using Max (T) permutation procedure implemented in PLINK software. The corrected values for multiple testing (corrected empirical *p*-values) were then calculated.

Cox proportional hazards model

The Cox proportional hazards model (3) or Cox regression model [43], is widely used in the analysis of time-to-event data [44-46].

$$h(t|x) = h_0(t)\exp(\beta_1 SNP_k + \beta_2 Sex + \beta_3 Alcohol + \beta_4 Smoking + \beta_5 BMI) \quad (3)$$

where $h(t|x)$ is the hazard at time t for a subject, $h_0(t)$ is the baseline hazard function. Then the hazard ratio (HR) is defined as the ratio of the predicated hazard function under two different values of a predictor variable. The PHREG procedure in SAS fits the Cox model by maximizing the partial likelihood function.

Parametric survival models

Several commonly used parametric distributions in survival models include exponential, Weibull, gamma, log-normal, and log-logistic [44-47].

$$\ln(T) = \beta_0 + \beta_1 SNP_k + \beta_2 Sex + \beta_3 Alcohol + \beta_4 Smoking + \beta_5 BMI + ln(\varepsilon) \quad (4)$$

where is *T* the time to event; $ln(\varepsilon)$ is the natural log of the error term. The exponentials of the β coefficients may be interpreted as the time ratio (TR) [45-46,48]. If TR >1, the event is less likely to occur; whereas if TR <1, the event is more likely to happen.

The LIFEREG procedure in SAS fits parametric survival models, where the link function can be taken from a class of distributions that include exponential, Weibull, log-normal, log-logistic, and gamma distributions.

Supremum test for proportional hazards assumption

Both the graphical and numerical methods [49] were used to check the proportional hazards assumption in the ASSESS option of PROC PHREG. These methods are based on cumulative sums of martingale residuals over follow-up times or covariate values. The ASSESS option plots the cumulative score residuals against time for each independent variable and the RESAMPLE option computes the *p*-value of a Kolmogorov-type supremum test based on a sample of 1,000 simulated residual patterns. A significant p-value indicates a poor fit. The parametric Weibull and the exponential regression models share the assumption of proportional hazards with the Cox regression model [50].

Evaluation criteria for goodness of fit

The Akaike information criterion (AIC) statistic [51-52] and the Bayesian information criterion (BIC) statistic [53] were used to measure the goodness of model fit and compare survival models.

$$AIC = -2\ln\{p(x|\hat{\theta})\}+2k \qquad (5)$$

and

$$BIC = -2\ln\{p(x|\hat{\theta})\}+k\ln n \qquad (6)$$

where x is the random variable, $\hat{\theta}$ is the maximum likelihood estimate, k is the number of parameters, and n is the sample size. Smaller AIC and/or BIC indicate a better model fit.

Survival analysis of AAO of T2D

The PHREG procedure in SAS was used to fit the Cox model; while the LIFEREG procedure was used to fit parametric survival models including the exponential, Weibull, log-normal, log-logistic, and gamma distributions. Multivariate Cox regression analysis and parametric survival analyses were conducted to detect associations of each SNP with AAO adjusting for gender, alcohol use in the past month, BMI and smoking status, respectively. The AIC and BIC values were used to compare the Cox regression and parametric survival models [45, 54-57]. Descriptive statistics, Cox regression, and parametric models analyses were conducted with SAS v.9.4 (SAS Institute, Cary, NC, USA). SAS codes are listed in Appendix.

Linkage disequilibrium and Haplotype block

To examine the relationships among the SNPs within the EXT2 gene, the pairwise linkage disequilibrium (LD) statistics (r^2) based on the HapMap data were calculated in the HAPLOVIEW software [58]. Haplotype blocks were built with stringent criteria, that SNPs within each block have strong LD with each other, sometimes resulting in splitting of visually recognized blocks.

Results

Descriptive statistics and genotype quality control

The demographic characteristics of the subjects 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 T2D ranged from 26 to 90 years. Two SNPs with MAF<5% were removed and all the left 20 SNPs were in Hardy-Weinberg equilibrium in the controls (p>0.05).

Table 1. Descriptive characteristics of cases and controls.

	Non-Diabetes	Type 2 Diabetes
Number Sex, N (%)	2686	878
Males	1051(39%)	424(48%)
Females	1635(61%)	454(52%)
BMI, kg/m^2		
Mean ± SD	28.8±5.2	32.4±6.7
Range	16.1-61.3	16.8-64.4
Alcohol, N (%)		
No	930(35%)	418(48%)
Yes	1752(65%)	456(52%)
Smoking, N (%)		
Never	1405(52%)	327(47%)
Current	245(9%)	54(7%)
Past	1032(39%)	331(46%)
Age, years		
Mean ± SD	65.4±11.4	69.2±10.6
Range	46-90	46-90
AAO, years		
Mean ± SD	-	62.6±11.8
Range	-	26-90

Table 2. SNPs associated with risk and/or AAO of T2D(p<0.05).

SNP	Position	Allele[a]	MAF[b]	HWE[c]	OR-Diabetes[d]	p-Diabetes[e]	EMP2[f]	β-AAO[g]	p-AAO[h]	EMP2[i]
rs7111879	44090717	G	0.45	0.698	0.85(0.75,0.95)	6.33E-3	0.124	1.22(0.10, 2.33)	3.26E-2	0.515
rs4237646	44093796	G	0.26	0.352	0.87(0.74,1.04)	0.12	0.931	3.18(1.64,4.72)	5.79E-5	3.00E-3
rs4755230	44140920	G	0.27	0.312	0.91(0.77,1.07)	0.238	0.995	3.19(1.71,4.67)	2.76E-5	1.00E-3

[a] Minor allele; [b] Minor allele frequency; [c] Hardy-Weinberg equilibrium test p-value; [d] Odds ratio for diabetes based on multiple logistic regression; [e] p-value based on logistic regression; [f] Corrected empirical p-value generated by 100,000 permutation tests using Max (T) permutation procedure implemented in PLINK; [g] Regression coefficient for AAO based on multiple linear regression; [h] p-value based on linear regression; [i] Corrected empirical p-value generated by 100,000 permutation tests using Max (T) permutation procedure implemented in PLINK.

Table 3. Results of the Cox regression and parametric models in survival analysis of AAO of T2D.

Models	AIC[a]	Rank	BIC[b]	Rank	AIC[c]	Rank	BIC[d]	Rank	AIC[e]	Rank	BIC[f]	Rank
Cox	100336.1	6	10067	6	10009.3	6	10056.7	6	10010.7	6	10044.1	6
Weibull	6659.8	2	6702.7	2	6648.5	2	6691.5	2	6641.5	2	6684.4	2
Exponential	8969.3	5	9007.5	5	8968.9	5	9007.2	5	8959.2	5	8997.3	5
Log-logistic	6723.2	4	6766.1	4	6711.5	3	6754.4	3	6703.9	3	6746.9	3
Log-normal	6722.9	3	6765.8	3	6712.9	4	6755.8	4	6705.9	4	6748.8	4
Gamma	6653.8	1	6701.3	1	66642.6	1	6690.3	1	6635.5	1	6683.2	1

[a] AIC for rs7111879 adjusted for sex, alcohol use, smoking status, and BMI; [b] BIC for rs7111879 adjusted for sex, alcohol use, smoking status, and BMI; [c] AIC for rs4237646 adjusted for sex, alcohol use, smoking status, and BMI; [d] BIC for rs4237646 adjusted for sex, alcohol use, smoking status, and BMI; [e] AIC for rs4755230 adjusted for sex, alcohol use, smoking status, and BMI; [f] BIC for rs4755230 adjusted for sex, alcohol use, smoking status, and BMI.

Multiple linear and logistic regression analyses using PLINK

We found that 1 SNP associated with risk of T2D (rs7111879 with p=6.33 x 10^{-3}) and 3 SNPs associated with AAO (rs7111879, rs42376464 and rs4755230 with p=3.26 x 10^{-2}, 5.79 x 10^{-5} and 2.76 x 10^{-5}, respectively) (Table 2). Interestingly, the same SNP rs7111879 showed associations with both the risk and AAO of T2D. However, the associations of rs7111879 with risk and AAO were not significant after a Bonferroni correction (p>2.27x10^{-3}) or multiple testing correction using a permutation test (corrected p>0.05). The results of other 2 AAO associated SNPs (rs42376464 and rs4755230) remained significant after a Bonferroni correction (p<2.27x10^{-3}) and multiple testing correction using a permutation test (corrected p=3.0x10^{-3} and 1.0x10^{-3}, respectively).

Comparison of Cox Regression and Parametric Models using PROC PHREG and PROC LIFEREG

Table 3 shows the comparisons through AIC and BIC for the 6 types of models of the 3 SNPs associated with AAO (p<0.05). Overall, gamma distribution demonstrated the best model fit for all 20 SNPs, followed by the Weibull distribution. For example, rs4755230, the gamma distribution has the smallest AIC (AIC=6635.5), and the AIC for Weibull distribution is slightly larger (AIC=6641.5). BIC also indicated that Gamma (BIC=6683.2) and Weibull (BIC=6684.4) distribution had a similar fit and outperformed the rest models.

Supremum test for proportional hazards assumption

Figures 1 and 2 display the observed standardized score process with 20 simulated realizations from the null distribution for rs4755230 AA and AG genotypes, respectively. The plots showed that the observed process is atypical compared to the simulated realizations and revealed proportional hazards for the two genotypes compared with GG. The Kolmogorov-type supremum test results based on 1,000 simulations for all the covariates are shown in Table 4. The proportional hazards assumption was valid for all the variables (p>0.05).

Survival analysis of AAO using Cox regression, gamma and Weibull models

The results based on the Cox regression and parametric survival analyses using gamma and Weibull models are presented in Table 5. All the HR values for 3 SNPs are larger than 1 while all the TR values are smaller than 1. For example, the genotype AA of rs4755230 has HR=1.63, which indicates that the participant with AA has a 63% higher hazard rate of AAO than participant with GG. The TR using Weibull model is 0.93, which indicates that the participant with AA has a shortened AAO by 7% compared to participant with GG. In addition, the mean AAO was approximately 5.7 years earlier in the individuals who had two major allele (AA) of rs4755230 (mean AAO=61.6 years) compared with those who were homozygous for the minor allele (GG) (mean AAO = 67.3 years.

Figure 1. Explore plot for checking proportional hazards assumption for rs4755230AA compared with rs4755230GG.

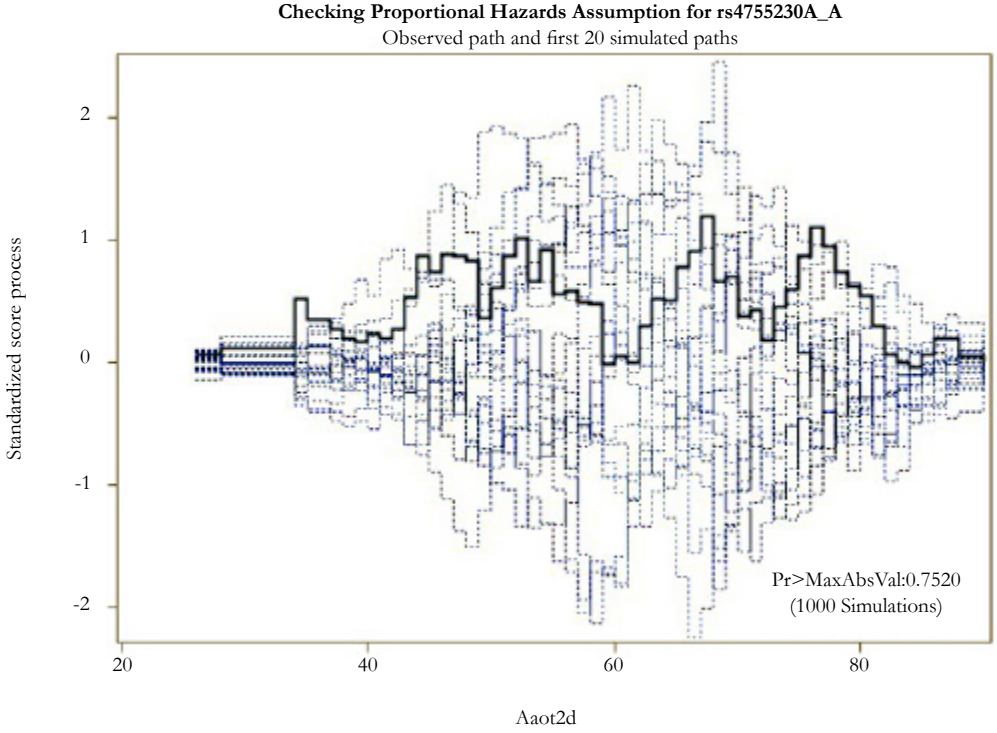
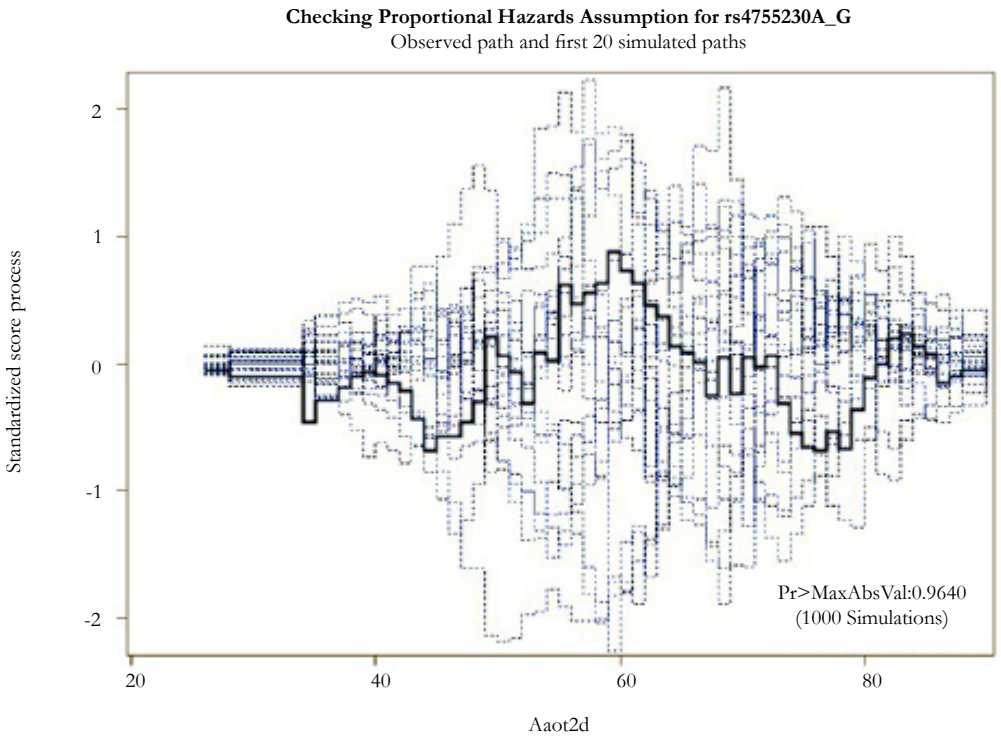


Figure 2. Explore plot for checking proportional hazards assumption for rs4755230AG compared with rs4755230GG.



Linkage disequilibrium and haplotype block

All SNPs including three previous identified T2D associated SNPs (rs3740878, rs1113132 and rs11037909) were within a haplotype block (Figure 3).

Discussion

To our knowledge, this is the first application to evaluate the Cox regression and 5 parametric survival models in genetic association analysis of the AAO of T2D. It is also the first candidate gene study to examine the associations of EXT2 gene polymorphisms with the AAO of T2D. In the procedure of model selection, the parametric gamma model outperformed the other models in the genetic association of the AAO of T2D, meanwhile we identified 1 SNP (rs7111879) associated with the risk of T2D and 3 SNPs (rs7111879, rs42376464 and rs4755230) associated with the AAO of T2D.

Table 4. Supremum Test for Proportional Hazards Assumption.

Variables	Maximum Absolute Value[a]	Replications	Seed	Pr > MaxAbsVal[b]
Sex	1.139	1000	1000	0.141
BMI	1.313	1000	1000	0.053
Alcohol	0.367	1000	1000	0.989
Smoking1	0.445	1000	1000	0.967
Smoking2	0.56	1000	1000	0.776
rs4755230AA	1.187	1000	1000	0.752
rs4755230AG	0.88	1000	1000	0.964

[a] Maximum absolute value based on the supremum test for proportional hazards assumption; [b]The p-value for the supremum test for proportional hazards assumption.

Table 5. Survival Analysis of 3 SNPs Associated with AAO using the Cox regression, gamma and Weibull models.

SNP	GT[a]	β(SE)[b]	p[c]	HR(95%CI)[d]	β(SE)[e]	p[f]	TR(95%CI)[g]	β(SE)[h]	p[i]	TR(95%CI)[j]
rs7111879										
	AA	0.206(0.103)	0.045	1.23(1.01,1.50)	-0.033(0.016)	0.038	0.97(0.94,0.99)	-0.036(0.017)	0.029	0.96(0.93,0.99)
	AG	0.118(0.095)	0.217	1.13(0.93,1.36)	-0.018(0.015)	0.225	0.98(0.95,1.01)	-0.018(0.015)	0.246	0.98(0.95,1.01)
	GG			1			1			1
rs42376464										
	TT	0.498(0.143)	0.0005	1.64(1.24,2.18)	-0.079(0.022)	0.0003	0.92(0.89,0.96)	-0.085(0.023)	0.0002	0.92(0.88,0.96)
	GT	0.374(0.146)	0.011	1.45(1.09,1.94)	-0.059(0.022)	0.008	0.94(0.90,0.98)	-0.064(0.024)	0.007	0.94(0.90,0.98)
	GG			1			1			1
rs4755230										
	AA	0.487(0.137)	0.0004	1.63(1.24,2.13)	-0.078(0.021)	0.0002	0.93(0.89,0.96)	-0.084(0.022)	0.0001	0.92(0.88,0.96)
	AG	0.386(0.140)	0.006	1.47(1.12,1.94)	-0.061(0.022)	0.005	0.94(0.90,0.98)	-0.066(0.023)	0.004	0.94(0.90,0.98)
	GG			1			1			1

[a] Genotype; [b] Regression coefficient and standard error (SE) based on the Cox regression; [c] p-value based on the Cox regression; [d] Hazard ratio (HR) and 95% confident interval (CI) based on the Cox regression; [e] Regression coefficient and SE based on the Weibull model; [f] p-value based on the Weibull model; [g]Time ratio (TR) and 95%CI based on the Weibull model; [h] Regression coefficient and SE based on the gamma distribution; [i] p-value based on the gamma model; [j] Time ratio (TR) and 95%CI based on the gamma model.

Cox regression and Weibull model have been extensively used to analyze incident diabetes [11-17] while non-parametric methods such as the Mann-Whitney and the Kruskall-Wallis tests [20], multiple linear regression [18] and Cox model [19] have been used to examine the associations of genetic variants with the AAO of T2D. However, no study was found to compare the Cox regression and parametric survival models in genetic association analysis of the AAO of T2D. Our results provided the first empirical evidence of the model comparisons in genetic studies of the AAO of T2D and showed that the parametric gamma and Weibull models performed better compared to Cox regression. It has been reported that the Weibull model shares the assumption of proportional hazards with the Cox regression model [50]. Particularly, if the assumption is met, then Weibull distribution provides an alternative, fully parametric approach to the Cox model, while if violated, other parametric models can be used with distributions rather than Weibull distribution [45]. In the present study, both the graphic and numeric methods in ASSESS statement in PROC PHREG showed that the assumption of proportional hazards is met. Therefore, Weibull model will be the first choice for genetic association study of AAO of T2D. In addition, the Weibull distribution provides similar HR estimates to Cox model; whereas a key strength of Weibull model allows the simultaneous estimates of treatment effects in terms of both HR and TR, which may lead to increase or decrease in survival time [59]. The survival function could be assumed in a certain form such as exponential, Weibull, and so on, with one or more parameters whose values are unknown, to be estimated from the real data [44]. Furthermore, if the shape of the survival distribution is known, parametric regression models may produce more efficient estimates than Cox model [60].

T2D is caused from the insulin resistance and β-cell dysfunction in the pancreas [61-63]. EXT2 gene is involved in the synthesis of heparin sulphate, abnormal bone growths (exostoses) [64] and in neural development [65]. The associations between three SNPs (rs3740878, rs1113132, rs11037909) in this gene and the risk of T2D have been previously reported in several studies [26, 35-36] but was not replicated in other studies [27-34]. Recently, the rs3740878 risk T allele was found to be nominally associated with reduced insulin secretion in carriers of the high-risk genotype compared with those with the low-risk genotype [29]; while two SNPs (rs3740878 and rs1113132) were associated with several measures of insulin resistance such as glucose and insulin levels in Pima Indians [66]. Furthermore, the EXT2 was found to have increased expression in brain, which suggested a possible site of action as to where this gene could affect diabetes risk [63]. A more

Figure 3. Linkage disequilibrium structure (r^2) within the EXT2 gene using the HapMap data.

recent study provided the first evidence on the relation between genetic defects in heparan sulfate synthesis and decreased pancreas anatomic volume with ensuing impaired beta-cell reserve capacity in carriers of loss-of-function mutations in EXT [67]. In the present study, we showed the first evidence of rs7111879 associated with the risk of T2D and 3 SNPs strongly associated with the AAO of T2D. Furthermore, the T2D associated SNP rs1113132 identified in the Bernelot Moens et al. study [67] also showed borderline association with the risk of T2D (p=0.0903) (data not shown). Additionally, three SNPs (rs3814767, rs7935138 and rs4379834) revealed nominal associations with the risk of T2D (p=0.0552, 0.0836 and 0.0835, respectively) (data not shown). However, the other two previously associated SNPs (rs3740878 and rs11037909) were not available in the Marshfield dataset. To examine the relationship among the SNPs within the EXT2 gene, we identified a haplotype block for 22 SNPs including rs3740878 and rs11037909 using the Hapmap data (Figure 3). We found that the three previously T2D risk associated SNPs (rs3740878, rs1113132, rs11037909) also had strong LD with the three nominal associated SNPs (rs3814767, rs7935138 and rs4379834 with r^2=0.78, 0.72, and 0.89, respectively) in the present study. Our results support a role of EXT2 in the development of T2D.

EXT2 is considered as a putative tumor suppressor gene and is associated with hereditary multiple exostoses, which is an autosomal dominant condition characterized by growth of multiple benign cartilage-capped tumors [22-23, 25, 68]. Recently, a gene expression study showed that the EXT2 gene activity in benign prostatic hyperplasia and prostate tumors was lower than that in normal prostate tissue [69]. Considering the comorbidity of T2D with multiple cancers such as endometrial and prostate cancers [4-8], it may be hypothesized that EXT2 gene may be involved in the pathogenesis of T2D and several cancers.

There are several strengths in this study. First, this study simultaneously demonstrated the performance of the semi-parametric Cox regression and five different parametric survival models in genetic association of the AAO of T2D with real data. Second, we examined 22 SNPs within the EXT2 gene and especially identified several genetic variants associated with the risk and AAO of T2D. Several limitations also need to be acknowledged. First, only one sample was used to examine the association of EXT2 gene with the risk of T2D due to limited data resources. Second, our current findings might be subject to type I error and need to be replicated in future studies.

Conclusions

The present study reveals that parametric gamma and Weibull models performed better than semi-parametric Cox proportional hazards model and other parametric models (including exponential, log-normal, and log-logistic) in the genetic association of the AAO of T2D. Furthermore, this is the first candidate gene study which investigates the associations between EXT2 SNPs and the AAO of T2D. These findings may serve as a resource for replication in other populations for future research on target genetic variation and the risk of T2D. Further functional study of the EXT2 gene will also help to better characterize the genetic basis of the risk and AAO of T2D.

Acknowledgments

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 RH00025, and Wisconsin Department of Commerce Technology Development Fund contract number TDF FYO10718. Funding support for genotyping, which was performed at Johns Hopkins University, was provided by the NIH (U01HG004438). 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

[1]. IDF (2009) IDF Diabetes Atlas. (4th edtn), Brussels.

[2]. Shin JK, Chiu YL, Choi S, Cho S, Bang H (2012) Serious psychological distress, health risk behaviors, and diabetes care among adults with type 2 diabetes: the California Health Interview Survey 2007. Diabetes Res Clin Pract 95(3): 406-414.

[3]. Halter JB, Musi N, McFarland Horne F, Crandall JP, et al. (2014) Diabetes and cardiovascular disease in older adults: current status and future directions. Diabetes 63(8): 2578-2589.

[4]. Inoue M, Iwasaki M, Otani T, Sasazuki S, Noda M, et al. (2006) Diabetes mellitus and the risk of cancer: results from a large-scale population-based cohort study in Japan. Arch Intern Med 166(17): 1871-1877.

[5]. Giovannucci E, Michaud D (2007) The role of obesity and related metabolic disturbances in cancers of the colon, prostate, and pancreas. Gastroenterology 132(6): 2208-2225.

[6]. Spurdle AB, Thompson DJ, Ahmed S, Ferguson K, Healey CS, et al. (2011) Genome-wide association study identifies a common variant associated with risk of endometrial cancer. Nat Genet 43(5): 451-454.

[7]. Bansal D, Bhansali A, Kapil G, Undela K, Tiwari P (2013) Type 2 diabetes and risk of prostate cancer: a meta-analysis of observational studies. Prostate Cancer Prostatic Dis 16(2): 151-158.

[8]. Wang KS, Owusu D, Pan Y, Xu C (2014) Common genetic variants in the HNF1B gene contribute to diabetes and multiple cancers. Austin Biomark Diagn 1(1): 5.

[9]. Elbers CC, Onland-Moret NC, Franke L, Niehoff AG, van der Schouw YT, et al. (2007) A Strategy to search for common obesity and type 2 diabetes genes. Trends Endocrinol Metab 18(1): 19-26.

[10]. Almgren P, Lehtovirta M, Isomaa B, Sarelin L, Taskinen MR, et al. (2011) Heritability and familiality of type 2 diabetes and related quantitative traits in the Botnia Study. Diabetologia 54(11): 2811-2819.

[11]. Bozorgmanesh M, Hadaegh F, Azizi F (2011) Predictive performance of the visceral adiposity index for a visceral adiposity-related risk: type 2 diabetes. Lipids Health Dis 10: 88.

[12]. Rosella LC, Manuel DG, Burchill C, Stukel TA; PHIAT-DM team (2011) A population-based risk algorithm for the development of diabetes: development and validation of the Diabetes Population Risk Tool (DPoRT). J Epidemiol Community Health 65(7): 613-620.

[13]. Bozorgmanesh M, Hadaegh F, Saadat N, Azizi F (2012) Fasting glucose cut-off point: where does the risk terminate? Tehran lipid and glucose study. Acta Diabetol 49(5): 341-348.

[14]. Adegbija O, Hoy W, Wang Z (2015) Predicting absolute risk of type 2 diabetes using age and waist circumference values in an aboriginal Australian community. PLoS One 10(4): e0123788.

[15]. Adegbija O, Hoy WE, Wang Z (2015) Corresponding waist circumference and body mass index values based on 10-year absolute type 2 diabetes risk in an Australian Aboriginal community. BMJ Open Diab Res Care 3(1): e000127.

[16]. Müller G, Wellmann J, Hartwig S, Greiser KH, Moebus S, et al. (2015) Association of neighbourhood unemployment rate with incident Type 2 diabetes mellitus in five German regions. Diabet Med 32(8): 1017-1022.

[17]. Gunderson EP, Hurston SR, Ning X, Lo JC, Crites Y, et al. (2015) Lactation and Progression to Type 2 Diabetes Mellitus After Gestational Diabetes Mellitus: A Prospective Cohort Study. Ann Intern Med 163(12): 889-898.

[18]. Papazoglou D, Papanas N, Papatheodorou K, Kotsiou S, Christakidis D, et al. (2006) An insertion/deletion polymorphism in the alpha2B adrenoceptor gene is associated with age at onset of type 2 diabetes mellitus. Exp Clin Endocrinol Diabetes 114(8): 424-427.

[19]. Lehman DM, Hunt KJ, Leach RJ, Hamlington J, Arya R, et al. (2007) Haplotypes of transcription factor 7-like 2 (TCF7L2) gene and its upstream region are associated with type 2 diabetes and age of onset in Mexican Americans. Diabetes 56(2): 389-393.

[20]. Bellanné-Chantelot C, Carette C, Riveline JP, Valéro R, Gautier JF, et al. (2008) The type and the position of HNF1A mutation modulate age at diagnosis of diabetes in patients with maturity-onset diabetes of the young (MODY)-3. Diabetes 57(2): 503-508.

[21]. Wu YQ, Heutink P, de Vries BB, Sandkuijl LA, van den Ouweland AM, et al. (1994) Assignment of a second locus for multiple exostoses to the pericentromeric region of chromosome 11. Hum Mol Genet 3(1): 167-171.

[22]. Hecht JT, Hogue D, Strong LC, Hansen MF, Blanton SH, et al. (1995) Hereditary multiple exostosis and chondrosarcoma: linkage to chromosome 11 and loss of heterozygosity for EXT-linked markers on chromosomes 11 and 8. Am J Hum Genet 56(5): 1125-1131.

[23]. Wuyts W, Ramlakhan S, Van Hul W, Hecht JT, van den Ouweland AM, et al. (1995) Refinement of the multiple exostoses locus (EXT2) to a 3-cM interval on chromosome 11. Am J Hum Genet 57(2): 382-387.

[24]. Blanton SH, Hogue D, Wagner M, Wells D, Young ID, et al. (1996) Hereditary multiple exostoses: confirmation of linkage to chromosomes 8 and 11. Am J Med Genet 62(2): 150-159.

[25]. Stickens D, Clines G, Burbee D, Ramos P, Thomas S, et al. (1996) The EXT2 multiple exostoses gene defines a family of putative tumour suppressor genes. Nat Genet 14(1): 25-32.

[26]. Sladek R, Rocheleau G, Rung J, Dina C, Shen L, et al. (2007) A genome-wide association study identifies novel risk loci for type 2 diabetes. Nature 445(7130): 881-885.

[27]. Omori S, Tanaka Y, Takahashi A, Hirose H, Kashiwagi A, et al. (2008) Association of CDKAL1, IGF2BP2, CDKN2A/B, HHEX, SLC30A8, and KCNJ11 with susceptibility to type 2 diabetes in a Japanese population. Diabetes 57(3): 791-795.

[28]. Lewis JP, Palmer ND, Hicks PJ, Sale MM, Langefeld CD, et al. (2008) Association analysis in African Americans of European-derived type 2 diabetes single nucleotide polymorphisms from whole-genome association studies. Diabetes 57(8): 2220-2225.

[29]. Moore AF, Jablonski KA, McAteer JB, Saxena R, Pollin TI, et al. (2008) Extension of type 2 diabetes genome-wide association scan results in the diabetes prevention program. Diabetes 57(9): 2503-2510.

[30]. Herder C, Rathmann W, Strassburger K, Finner H, Grallert H, et al. (2008) Variants of the PPARG, IGF2BP2, CDKAL1, HHEX, and TCF7L2 genes confer risk of type 2 diabetes independently of BMI in the German KORA studies. Horm Metab Res 40(10): 722-726.

[31]. Wu Y, Li H, Loos RJ, Yu Z, Ye X, et al. (2008) Common variants in CDKAL1, CDKN2A/B, IGF2BP2, SLC30A8, and HHEX/IDE genes are associated with type 2 diabetes and impaired fasting glucose in a Chinese Han population. Diabetes 57(10): 2834-2842.

[32]. Ren Q, Xiao J, Han X, Yang W, Ji L (2015) Impact of variants of the EXT2 gene on type 2 diabetes and its related traits in the Chinese han population. Endocr Res 40(2): 79-82.

[33]. Gutiérrez-Vidal R, Rodríguez-Trejo A, Canizales-Quinteros S, Herrera-Cornejo M, Granados-Silvestre MA, et al. (2011) LOC387761 polymorphism is associated with type 2 diabetes in the Mexican population. Genet Test Mol Biomarkers 15(1-2): 79-83.

[34]. Nemr R, Al-Busaidi AS, Sater MS, Echtay A, Saldanha FL, et al. (2013) Lack of replication of common EXT2 gene variants with susceptibility to type 2 diabetes in Lebanese Arabs. Diabetes Metab 39(6): 532-536.

[35]. Liu L, Yang X, Wang H, Cui G, Xu Y, et al. (2013) Association between variants of EXT2 and type 2 diabetes: a replication and meta-analysis. Hum Genet 132(2): 139-145.

[36]. Chang YC, Liu PH, Yu YH, Kuo SS, Chang TJ, et al (2014) Validation of type 2 diabetes risk variants identified by genome-wide association studies in Han Chinese population: a replication study and meta-analysis. PLoS One 9(4): e95045.

[37]. McCarty CA, Peissig P, Caldwell MD, Wilke RA (2008) The Marshfield Clinic Personalized Medicine Research Project: 2008 scientific update and lessons learned in the first 6 years. Personalized Medicine 5(5): 529-541.

[38]. McCarty CA, Wilke RA, Giampietro PF, Wesbrook SD, Caldwell MD (2005) Marshfield Clinic Personalized Medicine Research Project (PMRP): design, methods and recruitment for a large population-based biobank. Personalized Medicine 2(1): 49-79.
[39]. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, et al. (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 81(3): 559-575.
[40]. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, et al. (2006) Principal components analysis corrects for stratification in genome-wide association studies. Nature Genetics 38: 904-909.
[41]. Wang KS, Liu X, Zheng S, Zeng M, Pan Y, et al. (2012) A novel locus for body mass index on 5p15.2: a meta-analysis of two genome-wide association studies. Gene 500(1): 80-84.
[42]. Dunn OJ (1961) Multiple Comparisons Among Means. Journal of the American Statistical Association 56(293): 52-64.
[43]. Cox DR (1972) Regression models and life-tables. Journal of the Royal Statistical Society. Series B (Methodological) 34(2): 187-220.
[44]. Cantor AB (2003) SAS Survival Analysis Techniques for Medical Research. Springer Sciences and SAS Institute, Cary, USA.
[45]. George B, Seals S, Aban I (2014) Survival analysis and regression models. J Nucl Cardiol 21(4): 686-694.
[46]. Kasza J, Wraith D, Lamb K, Wolfe R (2014) Survival analysis of time-to-event data in respiratory health research studies. Respirology 19(4): 483-492.
[47]. Klein JP, Moeschberger ML (2003) Survival analysis: Techniques for censored and truncated data. Springer, New York.
[48]. Hernán MA, Cole SR, Margolick J, Cohen M, Robins JM (2005) Structural accelerated failure time models for survival analysis in studies with time-varying treatments. Pharmacoepidemiol Drug Saf 14(7): 477-491.
[49]. Lin DY, Wei LJ, Ying Z (1993) Checking the Cox Model with Cumulative Sums of Martingale-Based Residuals. Biometrika 80(3): 557-572.
[50]. Lee ET, Go OT (1997) Survival analysis in public health research. Annu Rev Public Health 18: 105-134.
[51]. Akaike H (1979) A Bayesian Extension of the Minimum AIC Procedure of Autoregressive Model Fitting. Biometrika 66(2): 237-242.
[52]. Akaike H (1981) Likelihood of a Model and Information Criteria. Journal of Econometrics 16(1): 3-14.
[53]. Simonoff JS (2003) Analyzing Categorical Data. Springer-Verlag, New York.
[54]. Malloy EJ, Spiegelman D, Eisen EA (2009) Comparing measures of model selection for penalized splines in Cox models. Comput Stat Data Anal 53(7): 2605-2616.
[55]. Wang SJ, Kalpathy-Cramer J, Kim JS, Fuller CD, Thomas CR (2010) Parametric survival models for predicting the benefit of adjuvant chemoradiotherapy in gallbladder cancer. AMIA Annu Symp Proc 2010: 847-851.
[56]. Ghadimi M, Mahmoodi M, Mohammad K, Zeraati H, Rasouli M, et al. (2011) Family history of the cancer on the survival of the patients with gastrointestinal cancer in northern Iran, using frailty models. BMC Gastroenterol 11: 104.
[57]. Zhu HP, Xia X, Yu CH, Adnan A, Liu SF, et al (2011) Application of Weibull model for survival of patients with gastric cancer. BMC Gastroenterol 11: 1.
[58]. Barrett JC, Fry B, Maller J, Daly MJ (2005) Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics 21(2): 263-265.
[59]. Carroll KJ (2003) On the use and utility of the Weibull model in the analysis of survival data. Control Clin Trials 24(6): 682-701.
[60]. Allison PD (2010) Survival Analysis Using SAS: A Practical Guide. SAS Institute, Cary, USA.
[61]. Weyer C, Bogardus C, Mott DM, Pratley RE (1999) The natural history of insulin secretory dysfunction and insulin resistance in the pathogenesis of type 2 diabetes mellitus. J Clin Invest 104(6): 787-794.
[62]. Staiger H, Machicao F, Stefan N, Tschritter O, Thamer C, et al. (2007) Polymorphisms within novel risk loci for type 2 diabetes determine beta-cell function. PLoS One 2(9): e832.
[63]. Ho MM, Yoganathan P, Chu KY, Karunakaran S, Johnson JD, et al. (2013) Diabetes genes identified by genome-wide association studies are regulated in mice by nutritional factors in metabolically relevant tissues and by glucose concentrations in islets. BMC Genet 14: 10.
[64]. Stickens D, Zak BM, Rougier N, Esko JD, Werb Z (2005) Mice deficient in Ext2 lack heparan sulfate and develop exostoses. Development 132(22): 5055-5068.
[65]. Inatani M, Yamaguchi Y (2003) Gene expression of EXT1 and EXT2 during mouse brain development. Brain Res Dev Brain Res 141(1–2): 129-136.
[66]. Rong R, Hanson RL, Ortiz D, Wiedrich C, Kobes S, et al. (2009) Association analysis of variation in/near FTO, CDKAL1, SLC30A8, HHEX, EXT2, IGF2BP2, LOC387761, and CDKN2B with type 2 diabetes and related quantitative traits in Pima Indians. Diabetes 58(2): 478-488.
[67]. Bernelot Moens SJ, Mooij HL, Hassing HC, Kruit JK, Witjes JJ, et al. (2014) Carriers of loss-of-function mutations in EXT display impaired pancreatic beta-cell reserve due to smaller pancreas volume. PLoS One 9(12): e115662.
[68]. Bridge JA, Nelson M, Orndal C, Bhatia P, Neff JR (1998) Clonal karyotypic abnormalities of the hereditary multiple exostoses chromosomal loci 8q24.1 (EXT1) and 11p11-12 (EXT2) in patients with sporadic and hereditary osteochondromas. Cancer 82(9): 1657-1663.
[69]. Suhovskih AV, Tsidulko AY, Kutsenko OS, Kovner AV, Aidagulova SV, et al. (2014) Transcriptional Activity of Heparan Sulfate Biosynthetic Machinery is Specifically Impaired in Benign Prostate Hyperplasia and Prostate Cancer. Front Oncol 4: 79.