Cancer Medicine

Open Access

SHORT REPORT

Significant SNPs have limited prediction ability for thyroid cancer

Shicheng Guo[1,2], Yu-Long Wang[3,4], Yi Li[1], Li Jin[1,5], Momiao Xiong[2], Qing-Hai Ji[3,4] & Jiucun Wang[1,5]

[1]State Key Laboratory of Genetic Engineering and Ministry of Education Key Laboratory of Contemporary Anthropology, School of Life Sciences, Fudan University, Shanghai 200433, China
[2]Human Genetics Center, University of Texas School of Public Health, Houston, Texas 77030
[3]Department of Head and Neck Surgery, Fudan University Shanghai Cancer Center, Shanghai 200032, China
[4]Department of Oncology, Shanghai Medical College, Fudan University, Shanghai 200032, China
[5]Fudan-Taizhou Institute of Health Sciences, Taizhou, Jiangsu 225300, China

Keywords
Genetic, risk prediction, SNPs, thyroid cancer

Correspondence
Jiucun Wang, National Ministry of Education Key Laboratory of Contemporary Anthropology, School of Life Sciences, Fudan University, Shanghai 200433, China.
Tel: +86-21-55665499; Fax: +86-21-556648845; E-mail: jcwang@fudan.edu.cn
Qing-Hai Ji, Department of Head& Neck Surgery, Fudan University Shanghai Cancer Center, 270 Dong'an Road, Shanghai 200032, China.
Tel: +86-21-64175590-2607;
Fax: +86-21-64174774;
E-mail: jiqinghai@shca.org.cn

Funding Information
This work was supported by the National Science Foundation of China (81001204 to Y.-L. Wang, 30872958 to Q.-H. Ji and 81270120 to J.-C. Wang), and grants from the Ministry of Science and Technology (2011BAI09B00 to L. Jin) and the Ministry of Health (201002007 to L. Jin).

Received: 30 October 2013; Revised: 21 December 2013; Accepted: 17 January 2014

***Cancer Medicine* 2014; 3(3): 731–735**

doi: 10.1002/cam4.211

Abstract

Recently, five thyroid cancer significantly associated genetic variants (rs965513, rs944289, rs116909374, rs966423, and rs2439302) have been discovered and validated in two independent GWAS and numerous case–control studies, which were conducted in different populations. We genotyped the above five single nucleotide polymorphisms (SNPs) in Han Chinese populations and performed thyroid cancer-risk predictions with nine machine learning methods. We found that four SNPs were significantly associated with thyroid cancer in Han Chinese population, while no polymorphism was observed for rs116909374. Small familial relative risks (1.02–1.05) and limited power to predict thyroid cancer (AUCs: 0.54–0.60) indicate limited clinical potential. Four significant SNPs have limited prediction ability for thyroid cancer.

Introduction

Thyroid cancer is the fifth most common type of female cancer and its incidence is increasing. It has been considered as one of highest familial risk carcinomas among all kinds of cancers [1, 2]. Most common diseases are caused by multiple genetic rather than few loci. In the last 2 years, two independent genome-wide association studies (GWAS) have been conducted to identify single nucleotide polymorphisms (SNPs) associated with thyroid cancer risk. Five SNPs (rs965513, rs944289, rs116909374, rs966423, and rs2439302) which were highly significantly associated with papillary thyroid carcinoma (PTC) were discovered by genome-wide association study. In addition, these five SNP were validated by continued case–control studies in more than three different populations (Han Chinese, Ohio, Poland, etc. Table 1).

To examine the prediction ability based on variants with highly significant associations, we use all five SNPs to predict thyroid cancer by nine classification methods (K-nearest neighbors, logistic regression, naïve Bayes, random forest, support vector machine, Bayesian additive regression trees (BART), recursive partitioning, fuzzy rule-based system, boosting). Contradictory to our intuitiveness, we found that although all these five SNPs were significantly associated with thyroid cancer, the precision of their prediction for thyroid cancer was very low.

Methods

The five SNPs were genotyped in 845 PTC and 1005 controls in Han Chinese population using the SNaPshot multiplex single-nucleotide extension system. PTC patients who were treated in the Department of Head and Neck Surgery, Fudan University Shanghai Cancer Center, Shanghai, China from January to December 2010 were enrolled in this study. All patients were ethnically Chinese Han and came from Eastern China. A total of 1005 cancer-free unrelated individuals were recruited from the Taizhou Longitudinal Study (TZL). The SNPs were genotyped with the SNaPshot multiplex single-nucleotide extension system. Details of SNPs (Table S1) and primers were listed in our previous article [3].

The relative risk to daughters of an affected thyroid cancer individual attributable to a given SNP is calculated by the formula: $\lambda^* = \frac{p(pr_2+qr_1)^2+q(pr_1+q)^2}{(p^2r_2+2pqr_1+q^2)^2}$, where p is the frequency of the risk allele, $q = 1 - p$, r_1 and r_2 are the relative risks (estimated by odds ratio [ORs]) for heterozygotes relative to common homozygotes and rare homozygotes relative to common homozygotes in the population, respectively [4, 5]. Assuming a multiplicative interaction, the proportion of the familial risk attributable to the SNP is calculated by $\log(\lambda^*)/\log(\lambda_o)$, where λ_o is the overall familial relative risk (FRR), estimated to be 8.48 for thyroid cancer [1]. Gender- and age-matched cases and controls were constructed by 1000 times resampling technology.

Nine machine learning methods were used to make prediction for PTC from health individuals, including K-nearest neighbors [6], logistic regression, naïve Bayes [7], random forest [8], support vector machine [7], BART [9], boosting, recursive partitioning, and fuzzy rule-based system [10]. The parameters in the models were optimally

Table 1. Odds ratio for five SNPs from GWAS and case–control association study in previous study.

			OR (*P*-value)^{12}					
Study	Population	Method	rs965513	rs944289	rs116909374	rs966423	rs2439302	Reference
1^{1}	Iceland	GWAS	1.73 (7.5e-13)	1.48 (8.6e-7)	–	–	–	[12]
	Iceland all	Combined	1.77 (6.8e-20)	1.44 (2.5e-8)				
	USA	Case–control	1.81 (1.2e-7)	1.32 (1.2e-2)				
	Spain	Case–control	1.54 (6.5e-3)	1.14 (4.3e-1)				
	USA and Spain	Case–control	1.72 (3.7e-9)	1.26 (1.1e-2)				
	All combined	Combined	1.75 (1.7e-27)	1.37 (2.0e-9)				
2^{1}	Chernobyl	GWAS	1.76 (4.9e-9)	1.13 (0.17)	–	–	–	[13]
		Combined	1.65 (4.8e-12)	–				
3^{2}	Japan	Case–control	1.69 (1.27e-4)	1.21 (0.0121)	–	–	–	[14]
4^{2}	UK	Case–control	1.98 (6.35e-34)	1.33 (6.95e-7)	–	–	–	[15]
5^{1}	Iceland	Case–control	1.70 (3.0e-18)	1.36 (4.2e-5)	2.03 (5.4e-7)	1.26(3.8e-4)	1.41 (1.3e-6)	[16]
	Netherland	Case–control	–	1.39 (0.013)	1.95 (0.024)	1.80(4.2e-6)	1.24 (0.088)	
	USA	Case–control	–	1.51 (0.0067)	1.98 (0.018)	1.36 (3.5e-3)	1.33 (6.1e-3)	
	Spain	Case–control	–	1.17 (0.31)	3.37 (2.6e-3)	1.20 (0.24)	1.34 (0.073)	
	All combined	Case–control	–	1.36 (4.9e-8)	2.09 (4.6e-11)	1.34 (1.3e-9)	1.36 (2.0e-9)	
6^{1}	USA	Case–control	2.10 (<2e-16)	1.28 (1.99e-3)	1.97 (1.11e-3)	1.35 (1.75e-4)	1.51 (4.24e-7)	[11]
	Poland	Case–control	1.78 (<2e-16)	1.21 (3.55e-3)	1.73 (6.27e-3)	1.15 (3.13e-2)	1.27 (2.20e-4)	
7^{12}	China	Case–control	1.53^{1} (7.1e-4) 1.53^{2} (1.4e-4)	1.51^{1} (2.8e-9) 1.53^{2} (2.0e-10)	–	1.32^{1} (0.006) 1.31^{2} (0.001)	1.40^{1} (2.1e-4) 1.41^{2} (2.7e-5)	[3]

GWAS, genome-wide association studies; OR, odds ratio.

[1] ORs were calculated based on the multiplicative model. For the combined study populations, the OR value were estimated using the Mantel–Haenszel model.

[2] ORs were calculated for the risk allele with using multiple logistic regression analyses.

Table 2. Estimation of familial relative risk of thyroid cancer for the five SNPs in population of Han Chinese.

SNPs	Familial relative risk	Proportion (100%)	*P*-value
rs965513	1.0189 (1.0186–1.0192)	0.843 (0.806–0.880)	<2.2e-16
s944289	1.0419 (1.0415–1.0422)	1.969 (1.922–2.016)	<2.2e-16
rs116909374	N.A.[1]	N.A.[1]	N.A.[1]
rs966423	1.0493 (1.0485–1.0500)	2.191 (2.093–2.289)	<2.2e-16
rs2439302	1.0207 (1.0205–1.0210)	0.977 (0.939–1.015)	<2.2e-16

[1] rs116909374 SNP was not detected in the Chinese population.

selected. Classification accuracy, sensitivity, specificity, and AUC were used to evaluate the performance of the methods. They were calculated by 10-fold cross-validation.

Results

Marginal FRR of the significant SNPs

As the previous studies showed that the five SNPs with large OR were significantly associated with thyroid cancer in various populations (Table 1). Our previous data also showed that SNPs were significantly associated with thyroid cancer in Chinese population (the seventh study of Table 1). In present study, we estimated the FRR for five significantly associated SNPs in Chinese population. We found that the FRRs were low, ranging from 1.02 to 1.05. These five SNPs counted only 5.98% of the overall familial risk (Table 2) which was very closed to that of polish population (about 6%) [11]. Our finding suggested that majority of the heritability was undiscovered.

Genetic risk prediction for thyroid cancer based on five SNPs

The five significant SNPs were used to predict risk of thyroid cancer by nine classification methods. The results were summarized in Table 3. The prediction accuracies ranged from 0.52 to 0.57 in the nine prediction methods, while receiver operating characteristics (ROCs) ranged from 0.54 to 0.60. The sensitivity of the prediction (0.28–0.48) was much less than specificity (0.56–0.76), which suggested the clinical application value might be limited (Table 3). In addition, the AUC of classification based on five SNPs and gender, and based on five SNPs, gender, and age ranged from 0.49 to 0.58, and from 0.50 to 0.59, respectively. This indicated that including gender and age information will not improve prediction (Tables S2 and S3, Fig. S1).

Conclusion

In the present study, we estimated the FRR and evaluated thyroid cancer prediction accuracy of the five SNPs that showed significant association with thyroid cancer in several association studies. The results showed that although the OR of each SNPs was large, the FRR of each SNPs was very marginal. By 10-fold cross-validation, we found that the prediction accuracy of five SNPs was low across all nine classification methods. Particularly, the sensitivity of five SNPs was very low. It suggested that the clinical application of five SNPs might be limited. Our results strongly demonstrate that complex diseases are caused by a large number of SNPs, environments, and their interactions. GWAS addressing common variants have come to its limit and missing heritability for most complex disorders is very high. Only about 5–10% heritability

Table 3. Model performance with methods based on five significant SNPs.

	AUC	Sensitivity	Specificity	Accuracy	Range of 95% CI of AUC
K-nearest neighbors	0.5589	0.3861	0.6591	0.533	[0.4293, 0.7101]
Logistic regression	0.6044	0.4982	0.5648	0.5346	[0.4433, 0.7368]
Naïve Bayes	0.5996	0.3921	0.7206	0.5686	[0.4571, 0.7469]
Random forest	0.5743	0.3169	0.7558	0.5535	[0.4405, 0.7233]
Support vector machine	0.5494	0.2762	0.7775	0.547	[0.4187, 0.7086]
Bayesian additive regression trees	0.5906	0.4779	0.5571	0.5211	[0.4385, 0.7211]
Boosting	0.6024	0.4723	0.5544	0.5157	[0.4584, 0.7287]
Recursive partitioning	0.5871	0.4085	0.7218	0.5778	[0.3926, 0.7048]
Fuzzy rule-based system	0.5396	0.4931	0.5006	0.4968	[0.4115, 0.6710]

AUC, sensitivity, specificity, and accuracy were its mean value in 10-fold validations. Range of 95% CI of AUC represents the range of the 95% CI of AUC in 10-fold Cross-validation. SVM represents support vector machines and Kernel Methods.

was found based on common disease common variant (CDCV) model. To improve prediction of genetic variation for complex diseases, we need to incorporate more common and rare SNPs, copy number variations (CNVs), and nongenetic susceptibility factors, such as iodine intake, exposure to radiation in the classification analysis. Novel statistical methods for variable screening should be developed to optimally select SNPs and CNVs across the genome for disease risk prediction.

Acknowledgment

This work was supported by the National Science Foundation of China (81001204 to Y.-L. Wang, 30872958 to Q.-H. Ji and 81270120 to J.-C. Wang), and grants from the Ministry of Science and Technology (2011BAI09B00 to L. Jin) and the Ministry of Health (201002007 to L. Jin).

Conflict of Interest

None declared.

References

1. Goldgar, D. E., D. F. Easton, L. A. Cannon-Albright, and M. H. Skolnick. 1994. Systematic population-based assessment of cancer risk in first-degree relatives of cancer probands. J. Natl. Cancer Inst. 86:1600–1608.
2. Dong, C., and K. Hemminki. 2001. Modification of cancer risks in offspring by sibling and parental cancers from 2,112,616 nuclear families. Int. J. Cancer 92:144–150.
3. Wang, Y. L., S. H. Feng, S. C. Guo, W. J. Wei, D. S. Li, Y. Wang, et al. 2013. Confirmation of papillary thyroid cancer susceptibility loci identified by genome-wide association studies of chromosomes 14q13, 9q22, 2q35 and 8p12 in a Chinese population. J. Med. Genet. 50:689–695.
4. Andrieu, N., G. Launoy, R. Guillois, C. Ory-Paoletti, and M. Gignoux. 2003. Familial relative risk of colorectal cancer: a population-based study. Eur. J. Cancer 39:1904–1911.
5. Cox, A., A. M. Dunning, M. Garcia-Closas, S. Balasubramanian, M. W. Reed, K. A. Pooley, et al. 2007. A common coding variant in CASP8 is associated with breast cancer risk. Nat. Genet. 39:352–358.
6. Brian, R. 2013. Class: functions for classification. *In* R. Brian, ed. Package 'class', ed. Version 7.3-7, Various functions for classification.
7. Meyer, D., E. Dimitriadou, K. Hornik, A. Weingessel, F. Leisch, C. Chang C. Lin. 2012. e1071: Misc functions of the department of statistics (e1071), TU Wien. *In* M. David, ed. Package 'e1071', ed. Version 1.5-11, Functions for latent class analysis, short time Fourier transform, fuzzy clustering, support vector machines, shortest path computation, bagged clustering, naive Bayes classifier.
8. Leo, B., C. Adele, L. Andy, and W. Matthew. 2012. randomForest: Breiman and Cutler's random forests for classification and regression. *In* L. Andy, ed. Package 'randomForest', ed. Version 4.6-7, Classification and regression based on a forest of trees using random inputs.
9. Chipman, H., E. George, and R. McCulloch. 2010. BART: Bayesian additive regression trees. Ann. Appl. Stat. 4:266–298.
10. Lala, S. R., B. Christoph, H. Francisco, M. B. Jose. 2013. frbs:Fuzzy Rule-based Systems for Classification and Regression Tasks. *In* B. Christoph, ed. Package 'frbs', ed. Version 2.1-0, This package implements functionality and various algorithms to build and use fuzzy rule-based systems (FRBSs).
11. Liyanarachchi, S., A. Wojcicka, W. Li, M. Czetwertynska, E. Stachlewska, R. Nagy, et al. 2013. Cumulative risk impact of five genetic variants associated with papillary thyroid carcinoma. Thyroid 23:1532–1540.
12. Gudmundsson, J., P. Sulem, D. F. Gudbjartsson, J. G. Jonasson, A. Sigurdsson, J. T. Bergthorsson, et al. 2009. Common variants on 9q22.33 and 14q13.3 predispose to thyroid cancer in European populations. Nat. Genet. 41:460–464.
13. Takahashi, M., V. A. Saenko, T. I. Rogounovitch, T. Kawaguchi, V. M. Drozd, H. Takigawa-Imamura, et al. 2010. The FOXE1 locus is a major genetic determinant for radiation-related thyroid carcinoma in Chernobyl. Hum. Mol. Genet. 19:2516–2523.
14. Matsuse, M., M. Takahashi, N. Mitsutake, E. Nishihara, M. Hirokawa, T. Kawaguchi, et al. 2012. The FOXE1 and NKX2-1 loci are associated with susceptibility to papillary thyroid carcinoma in the Japanese population. J. Med. Genet. 48:645–648.
15. Jones, A. M., K. M. Howarth, L. Martin, M. Gorman, R. Mihai, L. Moss, et al. 2012. Thyroid cancer susceptibility polymorphisms: confirmation of loci on chromosomes 9q22 and 14q13, validation of a recessive 8q24 locus and failure to replicate a locus on 5q24. J. Med. Genet. 49:158–163.
16. Gudmundsson, J., P. Sulem, D. F. Gudbjartsson, J. G. Jonasson, G. Masson, H. He, et al. 2012. Discovery of common variants associated with low TSH levels and thyroid cancer risk. Nat. Genet. 44:319–322.

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Figure S1. ROC comparison among all the machine learning prediction methods. Nine machine learning method were used to make prediction for PTC from health individuals, including K-nearest neighbors (KNN), logistic regression (LR), naïve Bayes, random forest, support vector machine, Bayesian additive regression trees

(BART), boosting, recursive partitioning, fuzzy rule-based system. The parameters in the models were optimally selected. Classification accuracy, sensitivity, specificity and AUC were used to evaluate the performance of the methods. They were calculated by 10-fold cross-validation.

Table S1. Genomic information for five Acknowledged SNPs from GWAS.
Table S2. Model performance with methods based on five SNPs and gender.
Table S3. Model performance with methods based on five SNPs, gender, and age.