**Significant SNPs have limited prediction ability for thyroid cancer**

**Abstract**

**Background**:

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.

**Method**: We genotyped five SNPs in Han Chinese populations and performed thyroid cancer risk predictions with nine machine learning methods.

**Result**: We found that the four SNPs were significantly associated with thyroid cancer in Han Chinese population while no polymorphism was observed. Small familial relative risks (1.02-1.05) and limited power to predict thyroid cancer (AUCs: 0.54-0.60) indicate limited clinical potential.

**Conclusion**: Four significant SNPs have limited prediction ability for thyroid cancer.

**Background**

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](#_ENREF_1), [2](#_ENREF_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 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 3 different populations (Han Chinese, Ohio and 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, 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 papillary thyroid carcinoma (PTC) and 1,005 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, Cancer Hospital, Fudan University, 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 1,005 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 (**Supplementary Table S1**) and primers were listed in our previous article [3](#_ENREF_3).

The relative risk to daughters of an affected thyroid cancer individual attributable to a given SNP is calculated by the formula: , where is the frequency of the risk allele, , and are the relative risks (estimated by odds ratios) for heterozygotes relative to common homozygotes and rare homozygotes relative to common homozygotes in the population, respectively[4](#_ENREF_4), [5](#_ENREF_5). Assuming a multiplicative interaction, the proportion of the familial risk attributable to the SNP is calculated by log()/ log(), where is the overall familial relativerisk (FRR), estimated to be 8.48 for thyroid cancer[1](#_ENREF_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](#_ENREF_6), logistic regression, naïve Bayes[7](#_ENREF_7), random forest[8](#_ENREF_8), support vector machine[7](#_ENREF_7), Bayesian additive regression trees (BART)[9](#_ENREF_9), boosting, recursive partitioning, fuzzy rule-based system[10](#_ENREF_10). 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.

**Results**

**Marginal familial relative risk of the significant SNPs**

As the previous studies showed that the five SNPs with large odds ratio (OR) were significantly associated with thyroid cancer in various populations (Table 1). Our previous data also showed there SNPs were significantly associated with thyroid cancer in Chinese population (the seventh study of table 1). In present study, we estimated the familial relative risk for five significantly associated SNPs in Chinese population. We found that the familial relative risks 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](#_ENREF_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 ROCs were ranged from 0.54 to 0.60. The sensitivity of the prediction (from 0.28 to 0.48) was much less than specificity (from 0.56 to 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 (**Supplementary Table S2-3, Figure S1**).

**Conclusion**

In present study, we estimated the familial relative risk 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 odds ratio of each SNPs was large, the familial relative risk of each SNPs was very marginal. By ten-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 was found based on CDCV model. To improve prediction of genetic variation for complex diseases we need to incorporate more common and rare SNPs, CNVs, and non-genetic 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.

**Acknowledgement**

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).

**Reference**

1. Goldgar DE, Easton DF, Cannon-Albright LA, Skolnick MH. Systematic population-based assessment of cancer risk in first-degree relatives of cancer probands. Journal of the National Cancer Institute 1994;86:1600-8.

2. Dong C, Hemminki K. Modification of cancer risks in offspring by sibling and parental cancers from 2,112,616 nuclear families. International journal of cancer. Journal international du cancer 2001;92:144-50.

3. Wang YL, Feng SH, Guo SC, Wei WJ, Li DS, Wang Y, Wang X, Wang ZY, Ma YY, Jin L, Ji QH, Wang JC. 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 2013.

4. Andrieu N, Launoy G, Guillois R, Ory-Paoletti C, Gignoux M. Familial relative risk of colorectal cancer: a population-based study. Eur J Cancer 2003;39:1904-11.

5. Cox A, Dunning AM, Garcia-Closas M, Balasubramanian S, Reed MW, Pooley KA, Scollen S, Baynes C, Ponder BA, Chanock S, Lissowska J, Brinton L, et al. A common coding variant in CASP8 is associated with breast cancer risk. Nature genetics 2007;39:352-8.

6. Brian R. class: Functions for Classification In: Brian R, ed. Package ‘class’, ed. Version 7.3-7, 2013:Various functions for classification.

7. E D, K H, F L, D M, A W. e1071: Misc Functions of the Department of Statistics (e1071), TU Wien. In: David M, ed. Package 'e1071', ed. Version 1.5-11, 2012: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, Adele C, Andy L, Matthew W. randomForest: Breiman and Cutler's random forests for classification and regression. In: Andy L, ed. Package ‘randomForest’, ed. Version 4.6-7, 2012:Classification and regression based on a forest of trees using random inputs.

9. Chipman H, George E, McCulloch R. BART: BAYESIAN ADDITIVE REGRESSION TREES. Annals of Applied Statistics 2010;4:266-98.

10. Lala SR, Christoph B, Francisco H, Jose MB. frbs:Fuzzy Rule-based Systems for Classification and Regression Tasks. In: Christoph B, ed. Package ‘frbs’, ed. Version 2.1-0, 2013:This package implements functionality and various algorithms to build and use fuzzy rule-based systems (FRBSs).

11. Liyanarachchi S, Wojcicka A, Li W, Czetwertynska M, Stachlewska E, Nagy R, Hoag K, Wen B, Ploski R, Ringel MD, Kozlowicz-Gudzinska I, Gierlikowski W, et al. Cumulative Risk Impact of Five Genetic Variants Associated With Papillary Thyroid Carcinoma. Thyroid 2013.

12. Gudmundsson J, Sulem P, Gudbjartsson DF, Jonasson JG, Sigurdsson A, Bergthorsson JT, He H, Blondal T, Geller F, Jakobsdottir M, Magnusdottir DN, Matthiasdottir S, et al. Common variants on 9q22.33 and 14q13.3 predispose to thyroid cancer in European populations. Nat Genet 2009;41:460-4.

13. Takahashi M, Saenko VA, Rogounovitch TI, Kawaguchi T, Drozd VM, Takigawa-Imamura H, Akulevich NM, Ratanajaraya C, Mitsutake N, Takamura N, Danilova LI, Lushchik ML, et al. The FOXE1 locus is a major genetic determinant for radiation-related thyroid carcinoma in Chernobyl. Hum Mol Genet 2010;19:2516-23.

14. Matsuse M, Takahashi M, Mitsutake N, Nishihara E, Hirokawa M, Kawaguchi T, Rogounovitch T, Saenko V, Bychkov A, Suzuki K, Matsuo K, Tajima K, et al. The FOXE1 and NKX2-1 loci are associated with susceptibility to papillary thyroid carcinoma in the Japanese population. J Med Genet 2012;48:645-8.

15. Jones AM, Howarth KM, Martin L, Gorman M, Mihai R, Moss L, Auton A, Lemon C, Mehanna H, Mohan H, Clarke SE, Wadsley J, et al. 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 2012;49:158-63.

16. Gudmundsson J, Sulem P, Gudbjartsson DF, Jonasson JG, Masson G, He H, Jonasdottir A, Sigurdsson A, Stacey SN, Johannsdottir H, Helgadottir HT, Li W, et al. Discovery of common variants associated with low TSH levels and thyroid cancer risk. Nat Genet 2012;44:319-22.

Supplementary Figure

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

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Study | Population | Method | OR(p-value)a,b | | | | | Reference |
| rs965513 | rs944289 | rs116909374 | rs966423 | rs2439302 |
| 1a | Iceland  Iceland all  USA  Spain  USA and Spain  All combined | GWAS  Combined  Case-control  Case-control  Case-control  Combined | 1.73(7.5e-13)  1.77(6.8e-20)  1.81(1.2e-7)  1.54(6.5e-3)  1.72(3.7e-9)  1.75(1.7e-27) | 1.48(8.6e-7)  1.44(2.5e-8)  1.32(1.2e-2)  1.14(4.3e-1)  1.26(1.1e-2)  1.37(2.0e-9) | - | - | - | [12](#_ENREF_12) |
| 2a | Chernobyl | GWAS  Combined | 1.76(4.9e-9)  1.65(4.8e-12) | 1.13(0.17)  - | - | - | - | [13](#_ENREF_13) |
| 3b | Japan | Case-control | 1.69(1.27e-4) | 1.21(0.0121) | - | - | - | [14](#_ENREF_14) |
| 4b | UK | Case-control | 1.98(6.35e-34) | 1.33 (6.95e-7) | - | - | - | [15](#_ENREF_15) |
| 5a | Iceland  Netherland  USA  Spain  All combined | Case-control  Case-control  Case-control  Case-control  Case-control | 1.70(3.0e-18)  -  -  -  - | 1.36(4.2e-5)  1.39(0.013)  1.51(0.0067)  1.17(0.31)  1.36(4.9e-8) | 2.03(5.4e-7)  1.95(0.024)  1.98(0.018)  3.37(2.6e-3)  2.09(4.6e-11) | 1.26(3.8e-4)  1.80(4.2e-6)  1.36(3.5e-3)  1.20(0.24)  1.34(1.3e-9) | 1.41(1.3e-6)  1.24(0.088)  1.33(6.1e-3)  1.34(0.073)  1.36(2.0e-9) | [16](#_ENREF_16) |
| 6a | USA  Poland | Case-control  Case-control | 2.10(<2e-16)  1.78(<2e-16) | 1.28(1.99e-3)  1.21(3.55e-3) | 1.97(1.11e-3)  1.73(6.27e-3) | 1.35(1.75e-4)  1.15(3.13e-2) | 1.51(4.24e-7)  1.27(2.20e-4) | [11](#_ENREF_11) |
| 7a,b | China | Case-control | 1.53a(7.1e-4)  1.53b(1.4e-4) | 1.51a(2.8e-9)  1.53b(2.0e-10) | - | 1.32a(0.006)  1.31b(0.001) | 1.40a(2.1e-4)  1.41b(2.7e-5) | [3](#_ENREF_3) |

a ORs were calculated based on the multiplicative model. For the combined study populations, the OR value were estimated using the Mantel-Haenszel model.

b 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.$ | N.A.$ | N.A.$ |
| 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 |

$rs116909374 SNP was not detected in the Chinese population

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 represent support vector machines and Kernel Methods