Gene 518 (2013) 107–113

Contents lists available at SciVerse ScienceDirect

Gene

journal homepage: www.elsevier.com/locate/gene

A new method for post Genome-Wide Association Study (GWAS) analysis of colorectal cancer in Taiwan

Hwei-Ming Wang [a,1], Tzu-Hao Chang [b,1], Feng-Mao Lin [c,1], Te-Hsin Chao [a], Wei-Chih Huang [c], Chao Liang [c], Chao-Fang Chu [c], Chih-Min Chiu [c], Wei-Yun Wu [c], Ming-Cheng Chen [d], Chen-Tsung Weng [e], Shun-Long Weng [c,f,g,h,i], Feng-Fan Chiang [a,**], Hsien-Da Huang [c,f,*]

[a] *Division of Colorectal Surgery, Department of Surgery, Taichung Veterans General Hospital, Taichung, Taiwan*
[b] *Graduate Institute of Biomedical Informatics, Taipei Medical University, Taipei, Taiwan*
[c] *Institute of Bioinformatics and Systems Biology, National Chiao Tung University, Hsin-Chu 300, Taiwan*
[d] *Division of Colorectal Surgery, Department of Surgery, Taichung Veterans General Hospital Puli Branch, Taichung, Taiwan*
[e] *Health GeneTech Corporation, Taoyuan, Taiwan*
[f] *Department of Biological Science and Technology, National Chiao Tung University, Hsin-Chu 300, Taiwan*
[g] *Department of Obstetrics and Gynecology, Hsinchu Mackay Memorial Hospital, Hsinchu, Taiwan*
[h] *Mackay Medicine, Nursing and Management College, Taipei, Taiwan*
[i] *Department of Medicine, Mackay Medical College, New Taipei City, Taiwan*

ARTICLE INFO

Article history:
Accepted 27 November 2012
Available online 20 December 2012

Keywords:
Genome-Wide Association Studies
GWAS
Single-nucleotide polymorphism
SNP
Colorectal cancer
Risk prediction

ABSTRACT

Recently, single nucleotide polymorphisms (SNPs) located in specific loci or genes have been identified associated with susceptibility to colorectal cancer (CRC) in Genome-Wide Association Studies (GWAS). However, in different ethnicities and regions, the genetic variations and the environmental factors can widely vary. Therefore, here we propose a post-GWAS analysis method to investigate the CRC susceptibility SNPs in Taiwan by conducting a replication analysis and bioinformatics analysis. One hundred and forty-four significant SNPs from published GWAS results were collected by a literature survey, and two hundred and eighteen CRC samples and 385 normal samples were collected for post-GWAS analysis. Finally, twenty-six significant SNPs were identified and reported as associated with susceptibility to colorectal cancer, other cancers, obesity, and celiac disease in a previous GWAS study. Functional analysis results of 26 SNPs indicate that most biological processes identified are involved in regulating immune responses and apoptosis. In addition, an efficient prediction model was constructed by applying Jackknife feature selection and ANOVA testing. As compared to another risk prediction model of CRC for European Caucasians population, which performs 0.616 of AUC by using 54 SNPs, the proposed model shows good performance in predicting CRC risk within the Taiwanese population, i.e., 0.724 AUC by using 16 SNPs. We believe that the proposed risk prediction model is highly promising for predicting CRC risk within the Taiwanese population. In addition, the functional analysis results could be helpful to explore the potential associated regulatory mechanisms that may be involved in CRC development.

1. Introduction

Pervasive worldwide, colorectal cancer (CRC) causes more than 608,700 deaths globally. The global incidence rate in males and females ranked fourth and third among all cancers, respectively, with over one million new cancer cases in 2008 (Jemal et al., 2011). In Taiwan, CRC ranks among the five leading cancers, with its incidence rate among all cancers ranking second highest in both males and females (Ferlay et al.,2010; Taiwan Cancer Registry, 2012). In general, cancer is mainly caused by genetic and environmental factors (Aaltonen et al., 2007; Anand et al., 2008; Lichtenstein et al., 2000; Powell et al., 1992), with most cancers requiring both factors for the development of clinical grade cancer. Moreover, the complexity of causation is owing to genetically susceptible individuals becoming more sensitive to potentially carcinogenic agents.

Abbreviations: GWAS, Genome-wide association study; SNP, Single-nucleotide polymorphism; CRC, Colorectal Cancer; LD, Linkage disequilibrium; ACU, Area under curve; ROC, Receiver operating characteristic.

* Correspondence to: H.-D. Huang, Institute of Bioinformatics and Systems Biology, National Chiao Tung University, Hsin-Chu 300, Taiwan. Tel.: +886 3 5712121x56952; fax: +886 3 5739320.

** Correspondence to: F.-F. Chiang, Division of Colorectal Surgery, Department of Surgery, Taichung Veterans General Hospital, Taichung, Taiwan. Tel.: +886 4 23592525x5160; fax: +886 4 23595046.

E-mail addresses: whm@vghtc.gov.tw (H.-M. Wang), kevinchang@tmu.edu.tw (T.-H. Chang), kiralintw@gmail.com (F.-M. Lin), thcchao@vghtc.gov.tw (T.-H. Chao), gohk0958@msn.com (W.-C. Huang), gsuper3150631@gmail.com (C. Liang), jeffrey.cfchu@gmail.com (C.-F. Chu), citrate34@gmail.com (C.-M. Chiu), ukpiyo@gmail.com (W.-Y. Wu), claudiochen7@gmail.com (M.-C. Chen), david@genebook.com.tw (C.-T. Weng), a4467@ms7.mmh.org.tw (S.-L. Weng), hankel.chiang@gmail.com (F.-F. Chiang), bryan@mail.nctu.edu.tw (H.-D. Huang).

[1] These authors contributed equally to this work.

0378-1119/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.gene.2012.11.067

Genome-wide association studies (GWAS) have explored in recent how single nucleotide polymorphisms (SNPs) and diseases are related, with a general focusing mainly on identifying the disease susceptibility loci and several disease-associated SNPs, which are spread across genomes between patients and healthy individuals. Several SNPs located in specific loci or genes have been identified in recent years as high risk or low risk variations of CRC (Tenesa et al., 2008; Tomlinson et al., 2007; Yang et al., 2009). However, across different ethnicities and regions, genetic variations (e.g., allele frequency and linkage disequilibrium (LD)) and environmental factors (e.g., dietary patterns, alcohol and smoking) can widely vary (Weir et al., 2005), resulting in lack of consensus regarding CRC susceptibility SNPs. Therefore, we propose a post-GWAS analysis method (Fig. 1) to investigate the CRC susceptibility SNPs in Taiwan by conducting a replication analysis and bioinformatics analysis, including statistical analysis, risk prediction model construction and *in silico* functional analysis.

2. Methods

Fig. 1 illustrates the proposed post-GWAS analysis method. Based on a literature survey, numerous significant SNPs from previous GWAS of CRC, cancer and CRC-related disease are collected for custom panel design. Numerous Taiwanese samples are obtained from Taichung Veterans General (VGHTC) Hospital for the replication analysis. By establishing the most significant genetic model for each SNP, the identified CRC susceptibility SNPs could be more useful for constructing the risk prediction models of good performance. Finally, an efficient prediction model is also constructed by Jackknife feature selection and ANOVA testing. Furthermore, based on *in silico* functional analysis of the significant SNPs, this study elucidates the potential associated regulatory mechanisms possibly involved in CRC development. The details of each part of the proposed method are described below.

2.1. Selection of single nucleotide polymorphisms

A literature search of colorectal cancer GWAS was undertaken using the electronic database PubMed. The search was limited to entries on human studies from 1995 up to the end of Jan 2011. In addition, the GWAS-significant SNPs in other cancers (e.g., esophageal cancer and gastric cancer) and CRC related diseases (e.g., obesity, ulcerative colitis and Crohn's disease (Goldacre et al., 2008; Okabayashi et al., 2012; Walczak et al., 2012)) were also collected for the replication analysis. This study focused mainly on collecting SNPs from GWAS of Asian populations. For GWAS of non-Asian populations, only SNPs with a similar allele frequency of the Asian population in HapMap dataset were collected. Finally, 144 SNPs were collected, including 17 CRC-associated SNPs, 114 cancer-associated SNPs and 13 SNPs of CRC related disease for the post-GWAS analysis.

2.2. Samples

All samples were Taiwanese and obtained from Taichung Veterans General (VGHTC) Hospital, Taiwan. All CRC samples were registered based on diagnoses made by physicians at VGHTC Hospital, and had pathologically proven adenocarcinoma. Additionally, samples were collected and clinicopathological information from CRC samples and normal samples was obtained with informed consent and approval from the institutional review board of VGHTC. Finally, 224 CRC samples and 400 normal samples were collected for the post-GWAS analysis.

2.3. Custom panel design

The SNP genotyping panel for replication analysis was designed using Illumina VeraCode Genotyping assay. Additionally, data analysis of custom panel results (including quality control, raw data normalization, clustering and genotype calling) was performed using Illumina GenomeStudio. Moreover, the genotyping accuracy of this panel was

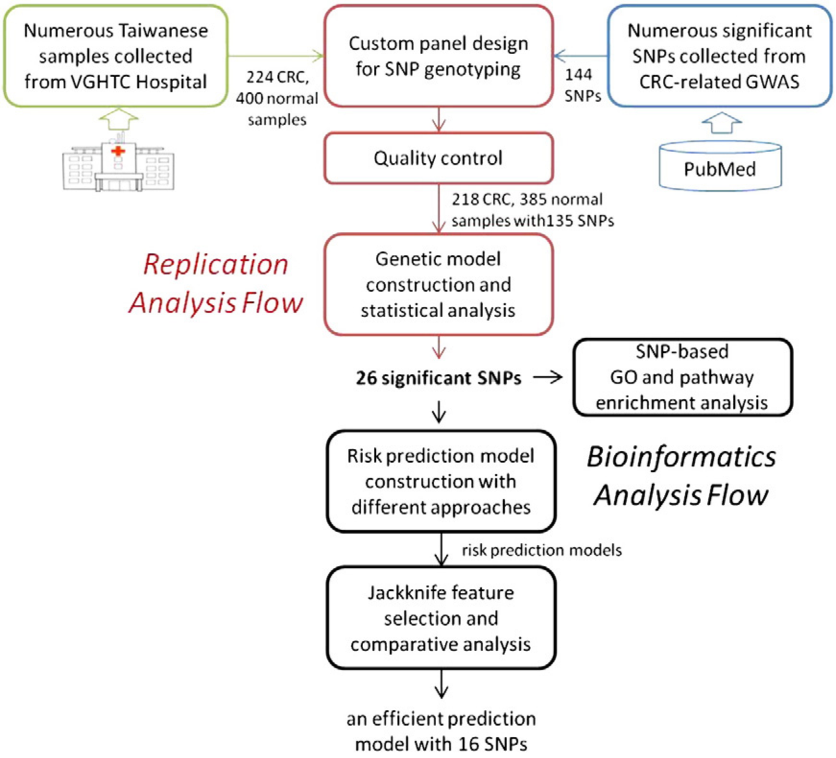
Numerous Taiwanese samples collected from VGHTC Hospital
224 CRC, 400 normal samples
Custom panel design for SNP genotyping
144 SNPs
Numerous significant SNPs collected from CRC-related GWAS
PubMed
Quality control
218 CRC, 385 normal samples with135 SNPs
Replication Analysis Flow
Genetic model construction and statistical analysis
26 significant SNPs
SNP-based GO and pathway enrichment analysis
Risk prediction model construction with different approaches
Bioinformatics Analysis Flow
risk prediction models
Jackknife feature selection and comparative analysis
an efficient prediction model with 16 SNPs

Fig. 1. The proposed post-GWAS analysis method of CRC.

validated using five duplicate samples with Illumina HumanOmni1-Quad BeadChip.

2.4. Data quality control

Samples with a call rate lower than 0.97 and SNPs with more than 2% missing data, Hardy Weinberg *P*-value smaller than 0.01, or minor allele frequency smaller than 0.05 were removed. Thus, statistical analysis consisted of 218 CRC samples, 385 normal samples and 135 SNPs selected from previous GWAS studies.

2.5. Genetic model construction and statistical analysis

This study also established a significant inheritance model for each SNP by calculating all possible combinations of genotypes and selecting the combination with the smallest deviance using logistic regression as its significantly well-fitted inheritance model. The missing value of an SNP was assigned to the category a sample distribution similar to that of the missing value in the inheritance model. Thus, significant SNPs of the replication analysis were identified, for use in constructing the risk prediction model.

2.6. Risk prediction model construction

Two approaches are developed to construct risk prediction models. Let ρ_i denote a binary indicator for the phenotypes of subjects $i = 1,\ldots,n$, and X_{ij} represent the dummy variable of SNP $j = 1,\ldots, m$ in inheritance model where m refers to significant SNPs identified in our replication analysis. In Approach 1, a linear logistic regression model is fitted as follows:

$$logit(P(\rho_i = 1|\chi_1)) = \beta_0 + \sum_{j=1}^{m} \beta_j X_{ij}$$

Thus, the personal odds ratio, denote as λ_i, is generated using the following formula:

$$\lambda_i = e^{\left(\sum_{j=1}^{m} \beta_j X_{ij}\right)}$$

Approach 2 considers an alternative method to calculate the personal odds ratio of subjects, denoted as γ_i. Based on single logistic regression analysis, γ_{ij} of each SNP is computed. Thus, γ_i is generated using the following formula:

$$\gamma_i = \prod\nolimits_1^m \gamma_{ij}$$

2.7. Feature selection and model evaluation

The prediction performance of each model was evaluated using 5-fold cross-validation with 100 sampling rounds and the area under receiver operating characteristic curve (AUC) (Kooperberg et al., 2010). The prediction sensitivity of models with different number of SNPs was then examined using the jackknife method to determine the priority of SNPs, iteratively. Next, a full model was constructed using the significant *p* SNPs, in which AUC of the full model was denoted as $\mathbf{AUC_{p:full}}$. Additionally, the reduced linear models constructed were generated by removing $\mathrm{SNP_i}$. Moreover, $\mathbf{\Delta AUC_i}$ was calculated using the following formula:

$$\Delta AUC_i = AUC_{p:full} - AUC_{i:reduced}, i = 1, \ldots, p$$

A smaller value of $\mathbf{\Delta AUC_i}$ implies a smaller advantage of $\mathrm{SNP_i}$ in contributing to $\mathbf{AUC_{p:full}}$. Thus, in this study, $\mathrm{SNP_i}$ with the smallest $\mathbf{\Delta AUC_i}$ was removed from the current set, with the remaining SNPs proceeding to the next round of feature selection until only two SNPs remained. Furthermore, the marginal significance of each SNP in the model was evaluated using the ANOVA test, subsequently allowing us to derive a compact model without marginally non-significant SNPs. Finally, the ANOVA test of SNPs and Likelihood test of models were performed using the R package *car* and *lmtest*, respectively.

3. Results

3.1. CRC susceptibility SNPs

Table 1 lists 26 significant SNPs ($P<0.05$) identified in the replication analysis, while Fig. 2 shows the CRC susceptibility loci. Five SNPs, rs231775 (Qi et al., 2010), rs1503185 (Mita et al., 2010), rs6983267 (Matsuo et al., 2009), rs7975232 (Mahmoudi et al., 2010) and rs10411210 (Houlston et al., 2008) were reported as associated with susceptibility to colorectal cancer in previous studies.

As a synonymous SNP of CTLA4, Rs231775 is an inhibitory receptor acting as a major negative regulator of T-cell responses (Thompson and Allison, 1997). Previous studies found that the genetic variations of CTLA4 interfere with gene activity (Anjos et al., 2002; Ligers et al., 2001), and were also associated with some immune-related diseases (Chistyakov et al., 2000). Because failure of the immune system may contribute to cancer (Mapara and Sykes, 2004), CTLA4 with altered functions may affect the development of CRC. As a non-synonymous SNP of PTPRJ, Rs1503185 inhibits cell growth by dephosphorylation of ERK1/2 in the MAPK signaling pathway mediated by various growth factors (Grazia Lampugnani et al., 2003; Palka et al., 2003). Interestingly, the polymorphism of PTPRJ may contribute to uncontrolled cell growth owing to reduced PTPRJ activity resulting from structural changes in the extracellular domain (Mita et al., 2010). As is well known, Rs6383267 is an intergenic SNP between POU5F1B and FAM84B. Previous studies have demonstrated that POU5F1B is a weak transcriptional activator possibly predisposed to carcinogenesis, including CRC (Haiman et al., 2007; Matsuo et al., 2009; Tomlinson et al., 2007; Wokolorczyk et al., 2008). FAM84B has been found to be expressed in esophageal squamous cell carcinomas (Huang et al., 2006). Rs7975232 is an intronic SNP of VDR. VDR encodes the nuclear hormone receptor for vitamin D3. As is widely assumed, VDR is involved in the carcinogenesis of CRC owing to its polymorphism which affects the serum levels of vitamin D. According to a previous study, CRC patients had more significantly lower levels of 25(OH) D than those with controls (Tangrea et al., 1997). As Rs10411210 is an intronic SNP of RHPN2, a previous study found a strong association with CRC in a meta-analysis of two GWA studies with a total of 13,315 individuals (Houlston et al., 2008).

Nineteen significant SNPs identified in our replication analysis were linked in previous GWAS studies with susceptibility to different cancers (Abnet et al., 2010; Bai et al., 2007; Bei et al., 2010; Hao et al., 2004; He et al., 2007; Hsiung et al., 2010; Ju et al., 2009, 2010; Long et al., 2010; Lu et al., 2009; Stacey et al., 2009; Sugimoto et al., 2007; Takata et al., 2010; Wang et al., 2006; Yuan et al., 2011). Also, several of their corresponding genes are related to oncogenesis, DNA repair, or apoptosis. Rs1412829 is an intronic SNP of CDKN2BAS. CDKN2BAS is located within the CDKN2B-CDKN2A gene cluster at chromosome 9p21 locus, which is linked to several cancers (Bei et al., 2010; Wrensch et al., 2009). Rs4072037 is a complex SNP of MUC1. MUC1 encodes a membrane-bound protein that belongs to the mucin family. As O-glycosylated proteins, mucins play an essential role in forming protective mucous barriers on epithelial surfaces. Previous studies have associated overexpression, aberrant intracellular localization, and changes in glycosylation of MUC1 with carcinomas (Bai et al., 2007; Yamamoto et al., 1997). Rs4430796 is an intronic SNP of HNF1B. In addition to encoding a member of the homeodomain-containing superfamily of transcription factors, HNF1B is associated with familial predisposition to prostate cancer. Moreover, its expression is altered in some cancers (Eeles et al., 2008). Rs1512268 is an intergenic SNP between NKX3-1 and

Table 1
26 CRC susceptibility SNPs in the replication analysis.

Loci	SNP ID	Chr Pos GRCh37.1 (bp)	Nearest gene(s)	SNP type	Cancer/ disease type	Major/ minor allele	*P*-value	Genetic model	Genotypic odds ratio	Ref.
1q22	rs4072037	155162067	MUC1	Complex	GC	G/A	4.19*E* − 02	GG/AG/AA	–/0.51/0.40	(Abnet et al., 2010)
1q32.1	rs1800872	206946407	IL19/IL10	Intergenic	GC	C/A	1.58*E* − 02	A-/CC	–/1.85	(Sugimoto et al., 2007)
2q33.2	rs231775	204732714	CTLA4	Synonymous	CRC	A/G	4.34*E* − 03	AA/G-	-/0.47	(Qi et al., 2010)
3p25.1	rs3731055	14220439	LSM3	Intron	LC	A/G	1.63*E* − 02	AA/G-	–/3.24	(Bai et al., 2007)
4q35.1	rs11721827	186991137	TLR3	Intron	NPC	C/A	4.38*E* − 02	AA/C-	–/1.46	(He et al., 2007)
5p15.33	rs2736100	1286516	TERT	Intron	LC	G/T	4.98*E* − 02	G-/TT	–/1.42	(Hsiung et al., 2010)
5q14.3	rs160277	82837631	VCAN	Synonymous	GC	A/C	8.25*E* − 03	AA/AC/CC	–/0.51/0.39	(Ju et al., 2010)
6p21.1	rs1983891	41536427	FOXP4	Intron	PC	T/C	9.17*E* − 03	CC/CT/TT	–/1.55/2.05	(Takata et al., 2010)
6p22.1	rs2860580	29906691	HCG4P6/HLA-A	Intergenic	NPC	T/C	2.99*E* − 02	TC/(TT + CC)	–/1.45	(Bei et al., 2010)
6q25.1	rs712221	152180241	ESR1	Intron	Ob	T/A	4.31*E* − 02	A-/TT	–/0.66	(Chen et al., 2009)
7q32.3	rs157935	130585553	FLJ43663/KLF14	Intergenic	BCC	G/T	2.96*E* − 02	G-/TT	–/0.67	(Stacey et al., 2009)
8p21.2	rs1512268	23526463	NKX3-1/SLC25A37	Intergenic	PC	A/G	2.11*E* − 02	G-/AA	–/2.04	(Takata et al., 2010)
8q22.1	rs3214050	95186382	CDH17	Non-synonymous	HCC	T/C	6.08*E* − 03	TT/CT/CC	–/1.18/0.39	(Wang et al., 2006)
8q24.21	rs6983267	128413305	POU5F1B/FAM84B	Intergenic	CRC	G/T	4.79*E* − 02	G-/TT	–/0.70	(Tomlinson et al., 2007)
9p21.3	rs1412829	22043926	CDKN2BAS	Intron	NPC	C/T	3.96*E* − 02	(CC + TT)/CT	–/1.54	(Bei et al., 2010)
11p11.2	rs1503185	48146622	PTPRJ	Non-synonymous	CRC	T/C	3.73*E* − 02	TT/C-	–/2.22	(Mita et al., 2010)
11q13.2	rs869736	67205462	PTPRCAP/CORO1B	Intergenic	GC	T/G	3.66*E* − 03	GT/TT/GG	–/0.41/1.21	(Ju et al., 2009)
11q23.1	rs1946518	112035458	IL18/TEX12	Intergenic	CD	G/T	2.71*E* − 02	T-/GG	–/1.56	(Brophy et al., 2010)
12q13.11	rs7975232	48238837	VDR	Intron	CRC	A/C	3.96*E* − 02	(AA + CC)/AC	–/1.42	(Mahmoudi et al., 2010)
13q12.12	rs1572072	24127210	SACS/TNFRSF19	Intergenic	NPC	T/G	4.36*E* − 03	TT/G-	–/0.43	(Bei et al., 2010)
14q11.2	rs1760944	20923149	OSGEP	UTR	LC	C/A	6.10*E* − 03	(AA + CC)/AC	–/1.59	(Lu et al., 2009)
16q12.1	rs4784227	52599188	TOX3/CHD9	Intergenic	BC	T/C	2.75*E* − 02	T-/CC	–/1.46	(Long et al., 2010)
17q12.1612	rs4430796	36098040	HNF1B	Intron	PC	G/A	4.32*E* − 02	A-/GG	–/0.52	(Takata et al., 2010)
17q12.1955	rs3135967	33313729	LIG3	Intron	EC	G/A	3.71*E* − 02	A-/GG	–/1.95	(Hao et al., 2004)
19q13.11	rs10411210	33532300	RHPN2	Intron	CRC	T/C	2.78*E* − 02	TT/CT/CC	–/2.77/3.47	(Houlston et al., 2008)
19q13.31	rs1799782	44057574	XRCC1	Synonymous	GC	T/C	1.43*E* − 02	T-/CC	–/1.52	(Yuan et al., 2011)

Abbreviations: CRC: colorectal cancer, NPC: nasopharyngeal carcinoma, GC: gastric cancer, BCC: basal cell carcinoma, LC: lung cancer, BC: breast cancer, PC: prostate cancer, HCC: hepatocellular carcinoma, EC: esophageal cancer, CD: celiac disease, Ob: obesity.

SLC25A37. As a transcription factor that functions as a negative regulator of epithelial cell growth in prostate tissue, NKX3-1 acts as a tumor suppressor to control prostate carcinogenesis (Zhang et al., 2010). Rs1983891 is an intronic SNP of FOXP4. FOXP4 belongs to subfamily P of the forkhead box (FOX) transcription factor family. Many members of the forkhead box gene family have roles in oncogenesis (Koo et al., 2012). Rs2736100 is an intronic SNP of TERT. TERT is a part of telomerase, which has reverse transcriptase activity. As a ribonucleoprotein polymerase, telomerase maintains telomere ends by adding the telomere repeat TTAGGG. Deregulation of telomerase expression in somatic cells may be involved in oncogenesis (Moriarty et al., 2005). Rs1799782 is a synonymous SNP of XRCC1, and rs3135967 is an intronic SNP of LIG3. XRCC1 and LIG3 are involved in defective DNA strand-break repair and sister chromatid exchange following treatment with ionizing radiation and alkylating agents. Variants of XRCC1 are related to several cancers (Hao et al., 2004; Moreno et al., 2006). Rs1572072 is an intergenic SNP between SACS and TNFRSF19. Capable of mediating the activation of JNK and NF-kappa-B, TNFRsF19 can induce apoptosis by a caspase-independent mechanism (Eby et al., 2000). Rs4784227 are intergenic SNPs between TOX3 and CHD9. As a transcriptional coactivator of the p300/CBP-mediated transcription complex, TOX3 protects against cell death by inducing anti-apoptotic and repressing pro-apoptotic transcripts (Dittmer et al., 2011).

3.2. CRC risk prediction model

Following data quality control processing, the genotyping data of 218 cases and 385 controls and 26 significant SNPs were collected to construct the risk prediction models using different approaches. Fig. 3 shows AUCs of two approaches with different numbers of SNPs considered, which are depicted as red and black lines. This figure also reveals AUCs of approaches evaluated using 5-fold cross-validation and training data, which are depicted as solid lines and dashed lines, respectively. Consequently a larger number of SNPs imply a higher prediction accuracy. When the number of used SNPs ranges from 7 to 18, AUCs of approach 1 exceed those of approach 2. However, approach 1 performs inferior to approach 2 when the number of used SNPs exceeds 22, although the difference is only slight. This study attempts to reduce the overfitting problem of logistic regression models with a large number of features in approach 1 by constructing a compact model without marginally non-significant SNPs through means of backward logistic regression. In this model, 16 marginally significant SNPs were selected (Table S1) by ANOVA test; in addition, *P*-value of the likelihood ratio test of this model is less than $2.2e-16$. Additionally, AUC of the compact model is 0.724 (blue circle in Fig. 3), which is slightly lower than 0.734 of the full model (as denoted by the blue rectangle in Fig. 3). Fig. 4 shows ROC curves of the full model and compact model, indicating that no obvious loss of prediction accuracy occurs between these two models. Our risk prediction model (AUC = 0.724), in which 16 SNPs are used, shows good performance as compared to a previous study (Park et al., 2012) (AUC = 0.616), which was constructed by 54 foreseeable SNPs within the European Caucasian population. Even when family history or epidemiological risk factors were considered, AUC of their model ranges only from 0.629 to 0.658. We thus believe that the proposed model is highly promising for predicting the risk of CRC in Taiwanese population.

3.3. In silico functional analysis of susceptibility SNPs

Gene ontology and pathway enrichment analysis of genes near 26 CRC susceptibility SNPs were performed using DAVID (Huang da et al., 2009a,b). According to Table 2, the major molecular functions were associated with DNA binding and transcription factor-related activities. HNF1B, NKX3-1, FOXP4, POU5F1B, VDR and ESR1 are either transcription factors or have transcription factor activities. ESR1 encodes an estrogen receptor (i.e. a ligand-activated transcription factor), and the estrogen receptors are involved in breast cancer and endometrial cancer (Gionet et al., 2009; Stoner et al., 2000). Most biological processes

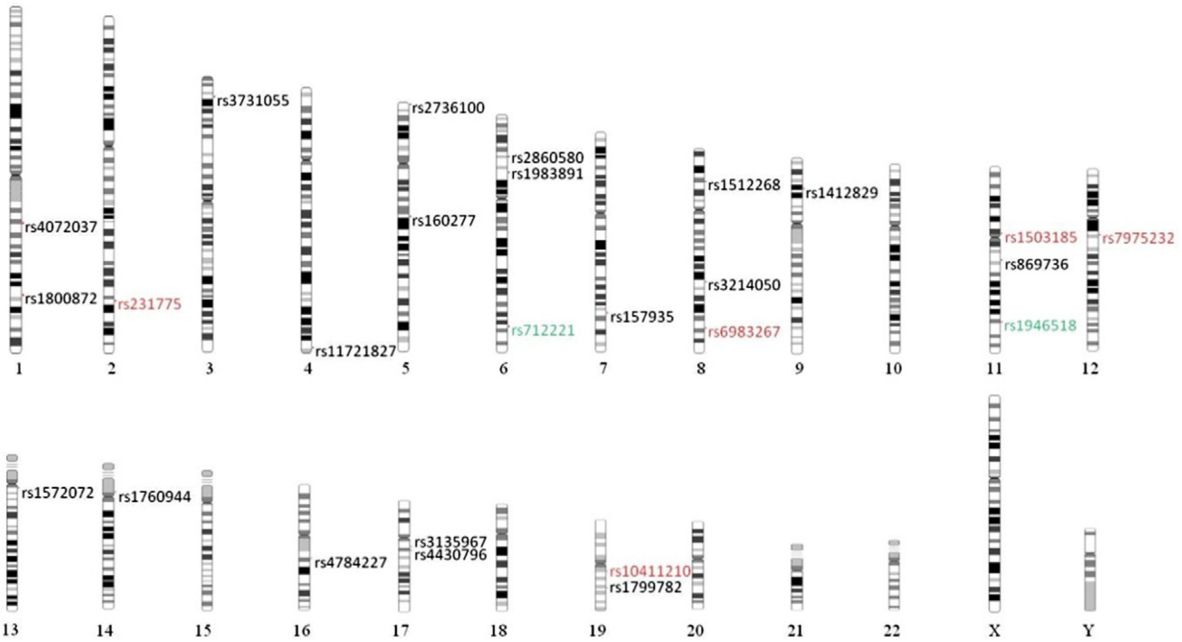

Fig. 2. CRC susceptibility loci identified in the replication analysis. Chromosome locations are based on chromosome build 37.1 GRCh37. The SNPs linked in previous GWAS studies with susceptibility to CRC, cancers and CRC-related diseases are color-coded with red, black and green, respectively.

identified are involved in regulating immune responses and apoptosis. TLR3, IL18, IL19, IL10, HLA-A and CTLA4 are the major genes involved in regulating the immune response. As a nucleotide-sensing TLR, TLR3 is activated by double-stranded RNA, a sign of viral infection (de Bouteiller et al., 2005). As a proinflammatory cytokine that augments natural killer cell activity in spleen cells, IL18 stimulates interferon gamma production in T-helper type I cells (Xu et al., 1998). IL19 may significantly contribute to inflammatory responses (Commins et al., 2008). IL10 has pleiotropic effects in immunoregulation and inflammation (Moore et al., 1993). HLA-A belongs to the heavy chain paralogues of HLA class I. Class I molecules play a major role in the immune system by presenting peptides derived from the endoplasmic reticulum lumen (Kalish, 1995). Genes contribute to the regulation of apoptosis are TERT, TNFRSF19, VDR, IL19, ESR1 and IL10. Notably, TNFRSF19, VDR, IL19 and IL10 participate in positive regulation of apoptosis. Additionally, NKX3-1, VDR, IL10 and CTLA4 are involved in the negative regulation of cell proliferation. According to Table 3, most identified pathways are related to immune responses, a phenomenon which is consistent with the results of GO enrichment. In particular, the base excision repair pathway should be noted, owing to that the genetic variations in base excision repair pathway genes are closely related to cancer risk (Hung et al., 2005); (Barry et al., 2011).

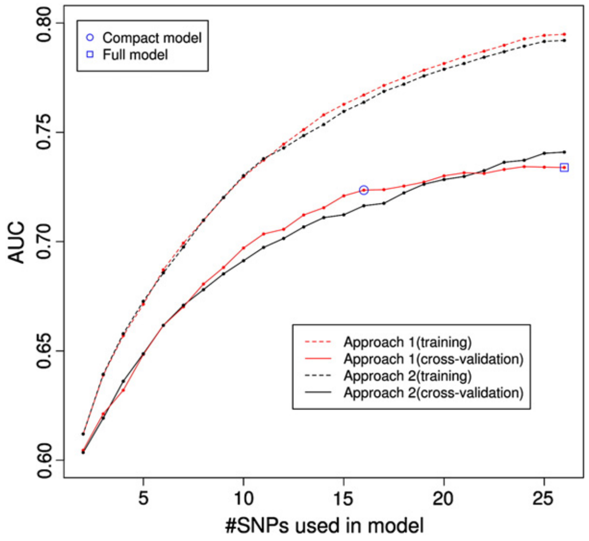

Fig. 3. AUC of CRC risk prediction models using different numbers of SNPs. AUCs of two approaches with different numbers of SNPs considered, which are depicted as red and black lines, and AUCs of approaches evaluated using 5-fold cross-validation and training data, which are depicted as solid lines and dashed lines, respectively. Additionally, AUC of the compact model is 0.724 (blue circle), which is slightly lower than 0.734 of the full model (blue rectangle).

4. Discussion and conclusions

By collecting the susceptibility SNPs of colorectal cancer in previous studies, this study attempts to identify the significant ones within the Taiwanese population by post-GWAS analysis. A custom panel was

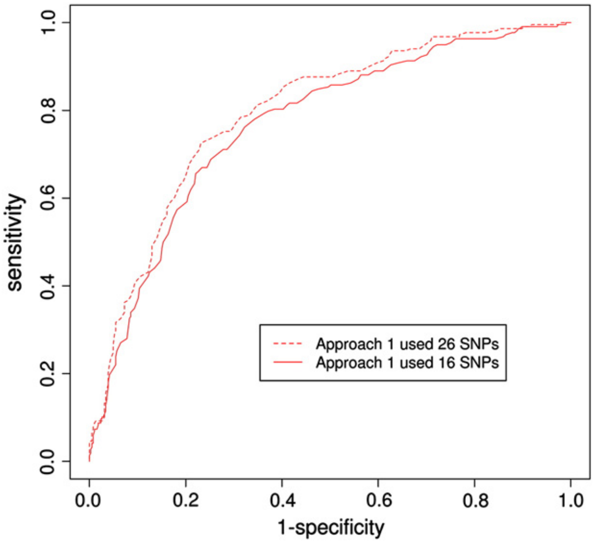

Fig. 4. ROC curve of the compact model and full mode.

Table 2
Results of SNP-based GO terms enrichment analysis.

Category	GO term	Term description	*P* value	Genes	FDR
GOTERM_BP_FAT	GO:0042129	regulation of T cell proliferation	0.006258	*IL18, IL10, CTLA4*	8.831
GOTERM_BP_FAT	GO:0042092	T-helper 2 type immune response	0.007666	*IL18, IL10*	10.716
GOTERM_BP_FAT	GO:0006955	immune response	0.009161	*TLR3, IL18, IL19, IL10, HLA-A, CTLA4*	12.676
GOTERM_BP_FAT	GO:0045354	regulation of interferon-alpha biosynthetic process	0.009574	*TLR3, IL10*	13.211
GOTERM_BP_FAT	GO:0050670	regulation of lymphocyte proliferation	0.010987	*IL18, IL10, CTLA4*	15.016
GOTERM_BP_FAT	GO:0032944	regulation of mononuclear cell proliferation	0.011241	*IL18, IL10, CTLA4*	15.338
GOTERM_BP_FAT	GO:0070663	regulation of leukocyte proliferation	0.011241	*IL18, IL10, CTLA4*	15.338
GOTERM_BP_FAT	GO:0032647	regulation of interferon-alpha production	0.015276	*TLR3, IL10*	20.287
GOTERM_BP_FAT	GO:0042981	regulation of apoptosis	0.016975	*TERT, TNFRSF19, VDR, IL19, ESR1, IL10*	22.288
GOTERM_BP_FAT	GO:0043067	regulation of programmed cell death	0.017652	*TERT, TNFRSF19, VDR, IL19, ESR1, IL10*	23.072
GOTERM_BP_FAT	GO:0010941	regulation of cell death	0.017910	*TERT, TNFRSF19, VDR, IL19, ESR1, IL10*	23.370
GOTERM_BP_FAT	GO:0050863	regulation of T cell activation	0.021052	*IL18, IL10, CTLA4*	26.903
GOTERM_BP_FAT	GO:0032479	regulation of type I interferon production	0.028459	*TLR3, IL10*	34.638
GOTERM_BP_FAT	GO:0042089	cytokine biosynthetic process	0.030328	*IL18, IL19*	36.466
GOTERM_BP_FAT	GO:0008285	negative regulation of cell proliferation	0.031114	*NKX3-1, VDR, IL10, CTLA4*	37.220
GOTERM_BP_FAT	GO:0042107	cytokine metabolic process	0.032194	*IL18, IL19*	38.243
GOTERM_BP_FAT	GO:0051249	regulation of lymphocyte activation	0.032546	*IL18, IL10, CTLA4*	38.573
GOTERM_BP_FAT	GO:0002694	regulation of leukocyte activation	0.040124	*IL18, IL10, CTLA4*	45.291
GOTERM_BP_FAT	GO:0050865	regulation of cell activation	0.044141	*IL18, IL10, CTLA4*	48.568
GOTERM_BP_FAT	GO:0001817	regulation of cytokine production	0.046901	*TLR3, IL18, IL10*	50.712
GOTERM_BP_FAT	GO:0010033	response to organic substance	0.047033	*HNF1B, TLR3, XRCC1, ESR1, IL10*	50.813
GOTERM_BP_FAT	GO:0045449	regulation of transcription	0.047437	*HNF1B, NKX3-1, FOXP4, TLR3, POU5F1B, VDR, KLF14, ESR1, IL10, CHD9*	51.119
GOTERM_BP_FAT	GO:0043065	positive regulation of apoptosis	0.048272	*TNFRSF19, VDR, IL19, IL10*	51.747
GOTERM_BP_FAT	GO:0043068	positive regulation of programmed cell death	0.049106	*TNFRSF19, VDR, IL19, IL10*	52.366
GOTERM_BP_FAT	GO:0010942	positive regulation of cell death	0.049667	*TNFRSF19, VDR, IL19, IL10*	52.777
GOTERM_CC_FAT	GO:0044454	nuclear chromosome part	0.023490	*TERT, TEX12, VDR*	22.743
GOTERM_CC_FAT	GO:0044427	chromosomal part	0.038434	*TERT, TEX12, VDR, CHD9*	34.652
GOTERM_CC_FAT	GO:0000228	nuclear chromosome	0.039583	*TERT, TEX12, VDR*	35.495
GOTERM_MF_FAT	GO:0043565	sequence-specific DNA binding	0.000842	*HNF1B, NKX3-1, TERT, FOXP4, POU5F1B, VDR, ESR1*	0.919
GOTERM_MF_FAT	GO:0003677	DNA binding	0.002286	*HNF1B, NKX3-1, TERT, FOXP4, TOX3, POU5F1B, VDR, XRCC1, KLF14, ESR1, LIG3, CHD9*	2.479
GOTERM_MF_FAT	GO:0003700	transcription factor activity	0.035619	*HNF1B, NKX3-1, FOXP4, POU5F1B, VDR, ESR1*	32.825

designed with 144 significant SNPs from published GWAS results. Also, 218 cases and 385 controls were collected for post-GWAS analysis. Twenty-six CRC susceptibility SNPs were identified, for use in constructing risk prediction models. Functional analysis results of 26 SNPs indicate that most biological processes identified are involved in regulating immune responses and apoptosis. Moreover, the major molecular functions are associated with DNA binding and transcription factor activities. Correspondingly, most identified pathways are related to immune responses. Inflammation is assumed to be the cause of CRC for a long time (Balkwill and Mantovani, 2001). Thus, altering the regulatory function in immune systems (e.g., IL10, IL18, IL19, TLR3 and CTLA4) by genetic variations likely increases the susceptibility to CRC. Finally, compared to another risk prediction model of CRC, which performs 0.616 of AUC by using 54 SNPs, the proposed model shows good performance in predicting CRC risk within the Taiwanese population, i.e. 0.724 AUC by using 16 SNPs. We believe that the proposed risk prediction model is highly promising for predicting CRC risk within the Taiwanese population.

Supplementary data related to this article can be foundonline at http://dx.doi.org/10.1016/j.gene.2012.11.067

Acknowledgments

The authors would like to thank the National Science Council of the Republic of China for financially supporting this research under contract no. NSC 98-2311-B-009-004-MY3, NSC 99-2627-B-009-003, NSC 101-2311-B-009-003-MY3 and NSC 100-2627-B-009-002. This work was supported in part by the UST-UCSD International Center of Excellence in Advanced Bio-engineering sponsored by the Taiwan

Table 3
Results of SNP-based pathway enrichment analysis.

Category	Term	*P* value	Genes	FDR	List Total
BIOCARTA	h_nkcells Pathway: Ras-independent pathway in NK cell-mediated cytotoxicity	0.00564	IL18, HLA-A, HCG4B	4.536	10
KEGG_PATHWAY	hsa05320: autoimmune thyroid disease	0.00717	IL10, HLA-A, CTLA4	5.670	14
KEGG_PATHWAY	hsa04060: cytokine–cytokine receptor interaction	0.02633	TNFRSF19, IL18, IL19, IL10	19.463	14
KEGG_PATHWAY	hsa04514: cell adhesion molecules (CAMs)	0.04326	VCAN, HLA-A, CTLA4	30.151	14
BIOCARTA	h_mhc Pathway: antigen processing and presentation	0.06108	HLA-A, HCG4B	40.361	10
KEGG_PATHWAY	hsa03410: base excision repair	0.08597	XRCC1, LIG3	51.777	14
KEGG_PATHWAY	hsa05330: allograft rejection	0.08833	IL10, HLA-A	52.775	14
BIOCARTA	h_ctlPathway: CTL mediated immune response against target cells	0.09036	HLA-A, HCG4B	54.006	10

National Science Council I-RiCE Program under grant number: NSC 100-2911-I-009-101, and Veterans General Hospitals and University System of Taiwan (VGHUST) Joint Research Program under grant number: VGHUST101-G5-1-1. This work was also partially supported by MOE ATU.

References

Aaltonen, L., et al., 2007. Explaining the familial colorectal cancer risk associated with mismatch repair (MMR)-deficient and MMR-stable tumors. Clin. Cancer Res. 13 (1), 356–361.

Abnet, C.C., et al., 2010. A shared susceptibility locus in PLCE1 at 10q23 for gastric adenocarcinoma and esophageal squamous cell carcinoma. Nat. Genet. 42 (9), 764–767.

Anand, P., et al., 2008. Cancer is a preventable disease that requires major lifestyle changes. Pharm. Res. 25 (9), 2097–2116.

Anjos, S., et al., 2002. A common autoimmunity predisposing signal peptide variant of the cytotoxic T-lymphocyte antigen 4 results in inefficient glycosylation of the susceptibility allele. J. Biol. Chem. 277 (48), 46478–46486.

Bai, Y., et al., 2007. Sequence variations in DNA repair gene XPC is associated with lung cancer risk in a Chinese population: a case-control study. BMC Cancer 7, 81.

Balkwill, F., Mantovani, A., 2001. Inflammation and cancer: back to Virchow? Lancet 357 (9255), 539–545.

Barry, K.H., et al., 2011. Genetic variation in base excision repair pathway genes, pesticide exposure, and prostate cancer risk. Environ. Health Perspect. 119 (12), 1726–1732.

Bei, J.X., et al., 2010. A genome-wide association study of nasopharyngeal carcinoma identifies three new susceptibility loci. Nat. Genet. 42 (7), 599–603.

Brophy, K., et al., 2010. Evaluation of 6 candidate genes on chromosome 11q23 for coeliac disease susceptibility: a case control study. BMC. Med. Genet. 11, 76.

Chen, H.H., Lee, W.J., Fann, C.S., Bouchard, C., Pan, W.H., 2009. Severe obesity is associated with novel single nucleotide polymorphisms of the ESR1 and PPARgamma locus in Han Chinese. Am. J. Clin. Nutr. 90, 255–262.

Chistyakov, D.A., et al., 2000. Complex association analysis of graves disease using a set of polymorphic markers. Mol. Genet. Metab. 70 (3), 214–218.

Commins, S., et al., 2008. The extended IL-10 superfamily: IL-10, IL-19, IL-20, IL-22, IL-24, IL-26, IL-28, and IL-29. J. Allergy Clin. Immunol. 121 (5), 1108–1111.

de Bouteiller, O., et al., 2005. Recognition of double-stranded RNA by human toll-like receptor 3 and downstream receptor signaling requires multimerization and an acidic pH. J. Biol. Chem. 280 (46), 38133–38145.

Dittmer, S., et al., 2011. TOX3 is a neuronal survival factor that induces transcription depending on the presence of CITED1 or phosphorylated CREB in the transcriptionally active complex. J. Cell Sci. 124 (Pt 2), 252–260.

Eby, M.T., et al., 2000. TAJ, a novel member of the tumor necrosis factor receptor family, activates the c-Jun N-terminal kinase pathway and mediates caspase-independent cell death. J. Biol. Chem. 275 (20), 15336–15342.

Eeles, R.A., et al., 2008. Multiple newly identified loci associated with prostate cancer susceptibility. Nat. Genet. 40 (3), 316–321.

Ferlay, J.S.H., Bray, F., Forman, D., Mathers, C., Parkin, D.M., 2010. GLOBOCAN 2008 v1.2, Cancer Incidence and Mortality Worldwide:IARC CancerBase No. 10. Lyon, France: International Agency for Research on Cancer. http://globocan.iarc.fr (accessed on 1/7/2012).

Gionet, N., et al., 2009. NF-kappaB and estrogen receptor alpha interactions: Differential function in estrogen receptor-negative and -positive hormone-independent breast cancer cells. J. Cell. Biochem. 107 (3), 448–459.

Goldacre, M.J., et al., 2008. Cancer in patients with ulcerative colitis, Crohn's disease and coeliac disease: record linkage study. Eur. J. Gastroenterol. Hepatol. 20 (4), 297–304.

Grazia Lampugnani, M., et al., 2003. Contact inhibition of VEGF-induced proliferation requires vascular endothelial cadherin, beta-catenin, and the phosphatase DEP-1/CD148. J. Cell Biol. 161 (4), 793–804.

Haiman, C.A., et al., 2007. A common genetic risk factor for colorectal and prostate cancer. Nat. Genet. 39 (8), 954–956.

Hao, B., et al., 2004. Identification of genetic variants in base excision repair pathway and their associations with risk of esophageal squamous cell carcinoma. Cancer Res. 64 (12), 4378–4384.

He, J.F., et al., 2007. Genetic polymorphisms of TLR3 are associated with Nasopharyngeal carcinoma risk in Cantonese population. BMC Cancer 7, 194.

Houlston, R.S., et al., 2008. Meta-analysis of genome-wide association data identifies four new susceptibility loci for colorectal cancer. Nat. Genet. 40 (12), 1426–1435.

Hsiung, C.A., et al., 2010. The 5p15.33 locus is associated with risk of lung adenocarcinoma in never-smoking females in Asia. PLoS Genet. 6 (8).

Huang da, W., et al., 2009a. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res. 37 (1), 1–13.

Huang da, W., et al., 2009b. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat. Protoc. 4 (1), 44–57.

Huang, X.P., et al., 2006. Negative implication of C-MYC as an amplification target in esophageal cancer. Cancer Genet. Cytogenet. 165 (1), 20–24.

Hung, R.J., et al., 2005. Genetic polymorphisms in the base excision repair pathway and cancer risk: a HuGE review. Am. J. Epidemiol. 162 (10), 925–942.

Jemal, A., et al., 2011. Global cancer statistics. CA Cancer J. Clin. 61 (2), 69–90.

Ju, H., et al., 2009. A regulatory polymorphism at position −309 in PTPRCAP is associated with susceptibility to diffuse-type gastric cancer and gene expression. Neoplasia 11 (12), 1340–1347.

Ju, H., et al., 2010. Genetic variants A1826H and D2937Y in GAG-beta domain of versican influence susceptibility to intestinal-type gastric cancer. J. Cancer Res. Clin. Oncol. 136 (2), 195–201.

Kalish, R.S., 1995. Antigen processing: the gateway to the immune response. J. Am. Acad. Dermatol. 32 (4), 640–652.

Koo, C.Y., et al., 2012. FOXM1: from cancer initiation to progression and treatment. Biochim. Biophys. Acta 1819 (1), 28–37.

Kooperberg, C., et al., 2010. Risk prediction using genome-wide association studies. Genet. Epidemiol. 34 (7), 643–652.

Lichtenstein, P., et al., 2000. Environmental and heritable factors in the causation of cancer—analyses of cohorts of twins from Sweden, Denmark, and Finland. N. Engl. J. Med. 343 (2), 78–85.

Ligers, A., et al., 2001. CTLA-4 gene expression is influenced by promoter and exon 1 polymorphisms. Genes Immun. 2 (3), 145–152.

Long, J., et al., 2010. Identification of a functional genetic variant at 16q12.1 for breast cancer risk: results from the Asia Breast Cancer Consortium. PLoS Genet. 6 (6), e1001002.

Lu, J., et al., 2009. Functional characterization of a promoter polymorphism in APE1/Ref-1 that contributes to reduced lung cancer susceptibility. FASEB J. 23 (10), 3459–3469.

Mahmoudi, T., et al., 2010. Vitamin D receptor gene ApaI polymorphism is associated with susceptibility to colorectal cancer. Dig. Dis. Sci. 55 (7), 2008–2013.

Mapara, M.Y., Sykes, M., 2004. Tolerance and cancer: mechanisms of tumor evasion and strategies for breaking tolerance. J. Clin. Oncol. 22 (6), 1136–1151.

Matsuo, K., et al., 2009. Association between an 8q24 locus and the risk of colorectal cancer in Japanese. BMC Cancer 9, 379.

Mita, Y., et al., 2010. Missense polymorphisms of PTPRJ and PTPN13 genes affect susceptibility to a variety of human cancers. J. Cancer Res. Clin. Oncol. 136 (2), 249–259.

Moore, K.W., et al., 1993. Interleukin-10. Annu. Rev. Immunol. 11, 165–190.

Moreno, V., et al., 2006. Polymorphisms in genes of nucleotide and base excision repair: risk and prognosis of colorectal cancer. Clin. Cancer Res. 12 (7 Pt 1), 2101–2108.

Moriarty, T.J., et al., 2005. An anchor site-type defect in human telomerase that disrupts telomere length maintenance and cellular immortalization. Mol. Biol. Cell 16 (7), 3152–3161.

Okabayashi, K., et al., 2012. Body mass index category as a risk factor for colorectal adenomas: a systematic review and meta-analysis. Am. J. Gastroenterol. 107, 1175–1185.

Palka, H.L., et al., 2003. Hepatocyte growth factor receptor tyrosine kinase met is a substrate of the receptor protein–tyrosine phosphatase DEP-1. J. Biol. Chem. 278 (8), 5728–5735.

Park, J.H., et al., 2012. Potential usefulness of single nucleotide polymorphisms to identify persons at high cancer risk: an evaluation of seven common cancers. J. Clin. Oncol. 30 (17), 2157–2162.

Powell, S.M., et al., 1992. APC mutations occur early during colorectal tumorigenesis. Nature 359 (6392), 235–237.

Qi, P., et al., 2010. CTLA-4+49A>G polymorphism is associated with the risk but not with the progression of colorectal cancer in Chinese. Int. J. Colorectal Dis. 25 (1), 39–45.

Stacey, S.N., et al., 2009. New common variants affecting susceptibility to basal cell carcinoma. Nat. Genet. 41 (8), 909–914.

Stoner, M., et al., 2000. Inhibition of vascular endothelial growth factor expression in HEC1A endometrial cancer cells through interactions of estrogen receptor alpha and Sp3 proteins. J. Biol. Chem. 275 (30), 22769–22779.

Sugimoto, M., et al., 2007. Effects of interleukin-10 gene polymorphism on the development of gastric cancer and peptic ulcer in Japanese subjects. J. Gastroenterol. Hepatol. 22 (9), 1443–1449.

Taiwan Cancer Registry, D.o.H., Executive Yuan, Taiwan, 2012. Cancer registry annual report in Taiwan.

Takata, R., et al., 2010. Genome-wide association study identifies five new susceptibility loci for prostate cancer in the Japanese population. Nat. Genet. 42 (9), 751–754.

Tangrea, J., et al., 1997. Serum levels of vitamin D metabolites and the subsequent risk of colon and rectal cancer in Finnish men. Cancer Causes Control 8 (4), 615–625.

Tenesa, A., et al., 2008. Genome-wide association scan identifies a colorectal cancer susceptibility locus on 11q23 and replicates risk loci at 8q24 and 18q21. Nat. Genet. 40 (5), 631–637.

Thompson, C.B., Allison, J.P., 1997. The emerging role of CTLA-4 as an immune attenuator. Immunity 7 (4), 445–450.

Tomlinson, I., et al., 2007. A genome-wide association scan of tag SNPs identifies a susceptibility variant for colorectal cancer at 8q24.21. Nat. Genet. 39 (8), 984–988.

Walczak, A., et al., 2012. The IL-8 and IL-13 gene polymorphisms in inflammatory bowel disease and colorectal cancer. DNA Cell Biol.

Wang, X.Q., et al., 2006. Liver intestine-cadherin (CDH17) haplotype is associated with increased risk of hepatocellular carcinoma. Clin. Cancer Res. 12 (17), 5248–5252.

Weir, B.S., et al., 2005. Measures of human population structure show heterogeneity among genomic regions. Genome Res. 15 (11), 1468–1476.

Wokolorczyk, D., et al., 2008. A range of cancers is associated with the rs6983267 marker on chromosome 8. Cancer Res. 68 (23), 9982–9986.

Wrensch, M., et al., 2009. Variants in the CDKN2B and RTEL1 regions are associated with high-grade glioma susceptibility. Nat. Genet. 41 (8), 905–908.

Xu, D., et al., 1998. Selective expression and functions of interleukin 18 receptor on T helper (Th) type 1 but not Th2 cells. J. Exp. Med. 188 (8), 1485–1492.

Yamamoto, M., et al., 1997. Interaction of the DF3/MUC1 breast carcinoma-associated antigen and beta-catenin in cell adhesion. J. Biol. Chem. 272 (19), 12492–12494.

Yang, H., et al., 2009. A novel polymorphism rs1329149 of CYP2E1 and a known polymorphism rs671 of ALDH2 of alcohol metabolizing enzymes are associated with colorectal cancer in a southwestern Chinese population. Cancer Epidemiol. Biomarkers Prev. 18 (9), 2522–2527.

Yuan, T., et al., 2011. Association of DNA repair gene XRCC1 and XPD polymorphisms with genetic susceptibility to gastric cancer in a Chinese population. Cancer Epidemiol. 35 (2), 170–174.

Zhang, P., et al., 2010. Gene expression profiles in the PC-3 human prostate cancer cells induced by NKX3.1. Mol. Biol. Rep. 37 (3), 1505–1512.