의학박사 학위논문

Identification of single nucleotide polymorphisms and copy number variations associated with colorectal cancer susceptibility in Korean population

한국인에서 대장암과 연관된 유전자 단일염기다형성과 복제수 변이의 확인

2015 년 8 월

서울대학교 대학원

의과학과 의과학 전공

박 창 호

한국인에서 대장암과 연관된 유전자 단일염기다형성과 복제수 변이의 확인

지도교수 김 종 일

이 논문을 의학박사 학위논문으로 제출함
2015년 4월

서울대학교 대학원
의과학과 의과학전공
박 창 호

박창호의 의학박사 학위논문을 인준함
2015년 7월

위 원 장 _____ (인)
부위원장 _____ (인)
위 원 _____ (인)
위 원 _____ (인)
위 원 _____ (인)

Identification of single nucleotide polymorphisms and copy number variations associated with colorectal cancer susceptibility in Korean population

by

Changho Park

A thesis submitted to the Department of Biomedical Science in partial fulfillment of the requirement of the Degree of Doctor of Philosophy in Medical Science at Seoul National University College of Medicine

July 2015

Approved by Thesis Committee:

Professor _____Chairman

Professor _____Vice chairman

Professor _____

Professor _____

Professor _____

ABSTRACT

Identification of single nucleotide polymorphisms and copy number variations associated with colorectal cancer susceptibility in Korean population

Changho Park

Major in Biomedical Science

Department of Biomedical Science

Seoul National University Graduate School

Introduction: Colorectal cancer (CRC) is one of the most common cancer worldwide. Many genome-wide association studies (GWAS) have been performed in the past few years in order to identify genetic variants contributing to CRC risk. Recent GWAS studies have identified several common single nucleotide polymorphisms (SNPs) associated with CRC risk in populations of European descent. Considering a significant racial and ethnic diversity in the human genome, it is doubtful whether the GWAS identified CRC susceptibility SNPs discovered in European populations would also be relevant in the Korean population. In addition, the incidence of CRC has lately increased significantly

in Asian population, especially in Korean population. However, CRC-susceptibility genetic variants in Korean population are not well known. In this study, we aimed to identify genetic variants associated with colorectal cancer risk in Korean population. In the first part of this study, we evaluated associations between CRC risk and five SNPs, which have previously been reported to be European CRC-susceptibility SNPs, in Korean population. Furthermore, we performed genome-wide screening to identify novel SNPs associated with CRC in Korean population. In the second part of this study, we examined the presence of copy number variations (CNVs). Several GWAS studies have identified more than 20 CRC susceptibility loci; yet, these studies have focused only on the analysis of SNPs, and these variants account for only a small fraction of heritability in CRC. Recent studies have reported that copy number variations (CNVs) are considered as important human genomic variants related to cancer predisposition. However, the contribution of CNVs to the CRC development and progression remains unclear. Therefore, we performed array comparative genomic hybridization to identify CNVs associated with CRC.

Methods: To investigate the racial and ethnic diversity of CRC-susceptibility genetic variants, we genotyped 5 established European CRC-susceptibility SNPs (rs3802842, rs4779584, rs4939827, rs6983267 and rs10795668) in 198 CRC cases and 329 controls in Korean population. Genotyping was conducted using TaqMan SNP Genotyping Assays. To find novel genetic variants using genome-wide screening in Korea, Illumina HumanHap 370K and 610K

BeadChips were performed on 45 and 60 CRC patients, respectively. The frequencies of genotypes in CRC patients were compared with those of Korean HapMap data. 8 candidate CRC-susceptibility SNPs were selected (rs12266240, rs10491619, rs10941887, rs1859915, rs727235, rs17051076, rs1599695 and rs902960). Subsequently, genotyping for replication was done in 189 CRC cases and 190 controls using TaqMan SNP Genotyping Assays. In the second part of this study, we performed array comparative genomic hybridization using the Agilent 180K microarray to detect CNVs associated with CRC in 36 patient and 47 control specimens. Using breakpoint PCR, we identified precise breakpoint location of *PKD1L2* copy number deletion region. We validated the association between *PKD1L2* copy variation and CRC risk in 1,874 cases and 2,088 controls using break point PCR.

Results: Of the five SNPs that were associated with CRC susceptibility in the European population, rs4939827 in *SMAD7* was significantly associated with a decreased risk of Korean CRC [age/gender-adjusted OR (95%CI), *P*]: additive model, 0.67 (0.47-0.95), *P*=.024; dominant model, 0.59 (0.39-0.91), *P*=.016]. Rs10795668 [additive model, 0.61 (0.39-0.95), *P*=.029] and rs4779584 [dominant model, 2.50 (1.16-5.39), *P*=.020] were associated with CRC risk in males and females, respectively. Interestingly, in subgroup analysis with the clinical data, we found that rs4939827 had a significant association in advanced CRC patients with stage III/IV status [additive model, 0.63 (0.43-0.93), *P*=.020], lymph node metastasis [additive model, 0.50 (0.61-0.82), *P*=.006], and distant metastasis [additive model, 0.22 (0.08-0.66), *P*=.007]. Among 8

candidate CRC susceptibility SNPs selected from genome-wide screening, rs17051076 was found to be significantly associated with an increased risk of microsatellite instability-high (MSI-H) CRC [age/gender-adjusted OR (95%CI): additive model, 4.25 (1.51-11.98, *P*=.006); dominant model, 3.52 (1.13-10.94) , *P*=.030] in the replication study. In the second part of the study, among the candidate CNVs identified in this experiment, only one variation, which involved *PKD1L2*, was located in a protein coding region and was therefore chosen as a final candidate CRC-associated CNV. Using breakpoint PCR, we confirmed the true breakpoint of the *PKD1L2* deletion region, and validated the association between *PKD1L2* variation and CRC risk in 1,874 cases and 2,088 controls (OR=1.44, 95% CI=1.04-1.98, P=.028). In addition, CN loss of *PKD1L2* is associated with increased CRC risk in patients with age younger than 50 years (OR=2.14, 95% CI 1.39-3.30, $P=5.8\times10^{-4}$). In subgroup analysis according to BMI (body mass index), we found that the CN loss of *PKD1L2* with BMI above or equal to 25 exhibited a significant increase in CRC risk (OR=2.29, 95% CI 1.29-4.05, *P*=.005). Furthermore, *PKD1L2* CN loss with BMI above or equal to 25 and age below 50 is associated with a remarkably increased risk of colorectal cancer (OR=5.24, 95% CI 2.36-11.64, $P= 4.8\times10^{-5}$). In addition, we found that *PKD1L2* variation in obese patients (BMI>=25) was associated with poor survival rate (*P*=.026).

Conclusions: Among the 5 European CRC-susceptibility SNPs, rs4939827, rs10795668 and rs4779584 may contribute to the risk of CRC in Korean population as well as in European populations. The new susceptibility SNP

rs17051076 could be associated with MSI-H CRC in Koreans. Also, the common copy number variation in *PKD1L2* is associated with cancer predisposition in CRC patients, and CN loss of *PKD1L2* with BMI below or equal to 25 and/or age below 50 exhibited a significant increased risk of CRC. In addition, in obese patients, *PKD1L2* variation was associated with poor survival rate.

* This work is published in Journal of Gastroenterology and Hepatology (1).

Keywords: single-nucleotide polymorphisms; copy number variations; colorectal cancer; genome-wide association studies; microsatellite instability-high; PKD1L2, body mass index

Student number: 2010-31168

CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

aCGH: array comparative genomic hybridization

BMI: body mass index

CNV: copy number variation

CRC: colorectal cancer

GWAS: genome wide association studies

LD: linkage disequilibrium

MSI: microsatellite instability

OR: odds ratio

PCR: polymerase chain reaction

SNP: single nucleotide polymorphism

CHAPTER 1

Colorectal Cancer-Susceptibility Single Nucleotide Polymorphisms in Koreans

INTRODUCTION

Colorectal cancer (CRC)

Worldwide, colorectal cancer (CRC) is the third most common cancer in males and the second in females (2). In Korea, the incidence of CRC has increased rapidly during the last decade and CRC was the second most common cancer in men and the third in women (3, 4), as well as first in both sexes in 2012 (http://globocan.iarc.fr/). Although environmental factors such as life style and diet are major risk factors for colorectal cancer (CRC), familial or early-onset CRC suggests a fundamental genetic contribution to CRC development (5-7). Familial adenomatous polyposis (FAP) and hereditary nonpolyposis colorectal cancer (HNPCC) caused by high-penetrance mutations in genes such as *APC* and mismatch repair genes account for less than 5% of all CRC cases (8-10). However, most of CRCs are sporadic cancer that is considered multi-genic complex disorder implicated by non-Mendelian genetic factors like common low-risk variants (11, 12).

Genome-wide association study (GWAS)

Genome-wide association studies (GWAS) are tested for association between the enormous number of single-nucleotide polymorphisms (SNPs) and common,

complex diseases (13). GWAS usually compare the frequencies of SNPs between two groups: people with the disease and people without the disease. If one variant is shown to be more frequent in patients than in those without the disease, then this variant is to be associated with the disease. Since the first GWAS was reported in age-related macular degeneration (AMD) in 2005 (14), huge GWAS (as of 2015, the GWAS catalog includes over 2100 publications) have been reported on various common complex disease (15). To identify genomic variants contributing to CRC risk, lots of GWAS have been performed and reported. Previous GWAS studies in populations of European ancestry have identified several CRC-susceptibility single nucleotide polymorphisms (SNPs) that confer a modest increase in CRC risk (16-22). However, the allele frequencies of SNPs are known to be different across racial and ethnic groups (23, 24). Therefore, it is unclear whether the European GWAS-identified CRC-susceptibility genetic variants are likewise relevant in Asian populations.

Colorectal cancer in Korean population

Over the past two decades, CRC incidence rates have increased dramatically in Asian countries including Japan, Hong Kong, Singapore, and South Korea (25). According to the GLOBOCAN 2012 data (http://globocan.iarc.fr, accessed on 16/01/2015), Korea has the highest rate of CRC in the world (45/100,000 people/year). This dramatic increase of CRC incidence coincides with the adoption of the Western

lifestyle, and this phenomenon can be explained by gene-environment interaction. Thus, understanding the genetic predisposition to develop CRC is extremely important in terms of inferring causality and mechanisms of colorectal carcinogenesis, as well as translating the findings into prevention of CRC. However, CRC-susceptibility genetic variants are yet to be known in Korean population.

Purpose and design of study

We have conducted this present study to explore CRC-susceptibility genetic variants among Koreans. First, we performed replication study by genotyping of 5 SNPs among European GWAS-identified CRC-susceptibility SNPs reported before in 2008 (Table 1). The genotyping for replication was conducted by TaqMan SNP Genotyping Assays in Korean CRC patients and Korean controls. Second, we conducted genome-wide screening to find novel CRC susceptibility genetic variants. The genotyping was done in Koran CRC patients using Illumina 370K and 610K BeadChips and the frequencies of genotype data was compared to that of Korean HapMap data. The replication study in Korean CRC patients and controls was conducted using TaqMan SNP Genotyping Assays.

Table1. Cancer susceptibility loci identified through GWAS in CRC before in 2008.

Locus	SNP	Date	Reference
8q24.21	rs10505477	2007-07-08	Zanke *et al.* (22)
8q24.21	**rs6983267**	**2007-07-08**	**Tomlinson *et al.* (20)**
18q21.1	**rs4939827**	**2007-10-14**	**Broderick *et al.* (16)**
8q23.3	rs16892766	2008-03-30	Tomlinson *et al.* (21)
10p14	**rs10795668**	**2008-03-30**	**Tomlinson *et al.* (21)**
8q24.21	**rs6983267**	**2008-03-30**	**Tomlinson *et al.* (21)**
15q13.3	**rs4779584**	**2008-03-30**	**Tomlinson *et al.* (21)**
18q21.1	**rs4939827**	**2008-03-30**	**Tomlinson *et al.* (21)**
18q21.1	**rs4939827**	**2008-03-30**	**Tenesa *et al.* (19)**
8q24.21	rs7014346	2008-03-30	Tenesa *et al.* (19)
11q23.1	**rs3802842**	**2008-03-30**	**Tenesa *et al.* (19)**
20p12.3	rs961253	2008-11-16	Houlston *et al.* (26)
14q22.2	rs4444235	2008-11-16	Houlston *et al.* (26)
19q13.11	rs10411210	2008-11-16	Houlston *et al.*(26)
16q22.1	rs9929218	2008-11-16	Houlston *et al* (26).

Bold characters indicate SNPs that genotyped in this study

GWAS, genome wide association studies; CRC, colorectal cancer; SNP, single nucleotide polymorphism

MATERIALS AND METHODS

Study population

All of the subjects that participated in this study were unrelated Koreans. Genomic DNA was extracted from peripheral blood by using QIAamp® DNA blood maxi kits according to the manufacturer's protocol (Qiagen, California, USA). DNA concentration was measured using Nanodrop ND-1000 spectrophotometer (Nano-drop Technologies, Delaware, USA). We conducted genetic association studies using genomic DNA obtained from the CRC patients diagnosed at Samsung Medical Center, Seoul, Korea, and healthy controls enrolled in the Sihwa-Banwol Environmental Health Cohort (SBEHC) (27). This cohort study initiated in 2005 and has been recruiting participants through health interviews and medical health examinations. In the two cities located on the west coast of South Korea, we used the Banwol (Ansan city) cohort as controls in this study. All included CRC patents had pathological confirmation and gave informed consent to donate the blood samples. Any patient who were not available clinical information and withdrawn the agreement of genetic studies were excluded. The mean age of the CRC patients was 62.1 ± 12.2 years, and 50.9 ± 8.7 years for the control group (Table 2). In the CRC cases, 120 (60.6%) were male, whereas 132 (40.1%) were male in the controls. The clinical characteristics of the enrolled 198 CRC patients were shown in table 3. Among the 198 CRC cases, 34 patients (17%) had early-onset CRC; 43 (22%) proximal CRC; and, 19 (10%) and 101 (51%) AJCC stage III or IV CRC, respectively. Lymph node and distant metastasis were noticed in 94 (47%) and 25 (14%) patients, respectively. This study

protocol was approved by the Institution Review Boards of Samsung Medical Center.

Microsatellite instability (MSI) was determined by five standard microsatellite markers, as recommended by the National Cancer Institute Workshop on MSI (the mononucleotide loci [BAT 25 and BAT 26], and the dinucleotide loci [D5S346, D17S250, and D2S123]). Polymerase chain reactions (PCR) were performed with an ABI PRISM 310 Genetic analyser (Applied Biosystems, Foster City, CA, USA) and Genescan software (Applied Biosystems) (28). MSI-High (MSI-H) was defined as a CRC that exhibited instability in two or more among the five markers. MSI-Low (MSI-L) was defined as a CRC that showed instability in one of five markers. Microsatellite stability (MSS) was defined as CRC that showed stability in all of the markers. MSI status was evaluated in 182 patients and 19 patients (10%) had MSI-H CRC. Among the total of 34 CRC patients younger than 50 years, MSI status was evaluated in 31 patients and seven patients (23%) revealed MSIH CRC. However, there could not be identified any Lynch syndrome patient.

Replication Study for the European CRC-susceptibility SNPs in Korean population

Among previous reported in European CRC-susceptibility SNPs, we selected the following 5 SNPs based on the previously reported paper (26), rs6983267 (8q24) (17, 19, 20), rs4779584 (15q13) (18), rs10795668 (10p14) (19), rs3802842 (11q23) (21) and rs4939827 (18q21) (16, 19) (Table 1). Genotyping was conducted using TaqMan

SNP Genotyping Assays (Applied Biosystems, San Francisco, California, USA) in 198 Korean CRC patients and 329 controls. The observed minor allele of rs4779584 was different from that in previous studies, and genotyping of rs4779584 was performed on 181 controls. All genotyped 5 SNPs had an average genotyping call rate of > 95% and a *p*-value of Hardy–Weinberg equilibrium of > 0.01 in the controls.

Genome-wide screening, SNPs selection, and genotyping

The genome-wide screening for CRC-susceptibility SNPs was done with the Illumina HumanHap 370k BeadChips (Illumina, San Diego, CA, USA) for 45 Korean CRC patients and the Illumina HumanHap 610k BeadChips (Illumina, San Diego, CA, USA) for 60 Korean CRC patients. The Illumina HumanHap 370k and 610k BeadChips contain 373397 tag SNPs and 592532 tag SNPs, respectively. The minor allele frequency (MAF) data of CRC patients in the Illumina HumanHap 370k and 610k BeadChips was compared to the MAF data of the Korean individuals offered by Korean HapMap Project. According to the difference of MAF between CRC patients and Korean HapMap, eight candidate CRC-susceptibility SNPs [rs12266240 (10p11.2), rs727235 (4q26), rs10491619 (9p21.2), rs17051076 (4q28), rs1599695 (9p21), rs1859915 (14q24.2), rs902960 (8p22), and rs10941887 (5p14)] were selected. A replication study was performed on 189 Korean CRC cases and 190 controls using genotyping by TaqMan SNP Genotyping Assays (Applied Biosystems, San Francisco, California, USA). All 8 genotyped candidate SNPs had an average genotyping call rate of > 95% and a *p*-value of Hardy–Weinberg equilibrium of >

0.01 in the controls.

Statistical analysis

Descriptive statistics were generated using frequencies and percentages for categorical variables and using means and standard deviations for continuous variables. Differences between categorical variables were analyzed using $\chi 2$ tests and Fisher's exact tests, as appropriate. For genetic association studies between CRC cases and healthy controls, additive, dominant, and recessive models were tested using logistic regression analysis. For the additive model, subjects were assigned a dummy variable of 0, 1, or 2, representing the number of minor alleles they had for that SNP. For the dominant model, subjects were coded 1 if they had at least one minor allele and 0 if they had no risk minor allele. For the recessive model, subjects were assigned 0 if they had two risk minor alleles and 1 if otherwise. The logistic regression analysis for genetic association studies was adjusted for age and gender to control for confounding by age and gender, because there was a significant difference in age and gender between both groups. Age and gender adjusted odds ratio (OR) and the corresponding 95% confidence interval (CI) were calculated. Genetic association studies were performed according to gender. In addition, to evaluate genetic association in clinically interesting situations, we performed subgroup analysis in patients with young-onset CRC (diagnosed <50 years of age), MSI-H CRC, proximal (cecum to the splenic flexure) and distal CRC (descending colon to rectum), CRC with lymph node and distant metastasis, American Joint Committee on Cancer (AJCC)

stage III or IV, and poorly differentiated CRC. A *P* value of less than 0.05 was considered to indicate statistical significance. Statistical analysis was performed using Statistical Package for the Social Sciences (version 19.0; Chicago, Illinois, USA).

Table 2. Characteristics of CRC patients and controls.

Variables		Cases (n=198)	Controls (n=329)	p value
Age (years ± SD)		62.1 ± 12.2	50. 9 ± 8.7	<0.001†
Gender	Males	120 (60.6%)	132 (40.1%)	
	Females	78 (39.4%)	197 (59.9%)	<0.001‡

†p-value was calculated by the t-test.

‡p value was calculated by the chi- square test.

Table3. Demographic and clinicopathological characteristics of 198 patients with colorectal cancer.

Variables		n (%)	Variables			n (%)
Age at diagnosis	≥ 50 y	164 (83)	TNM classification	T	1	18 (9)
	< 50 y	34 (17)			2	27 (14)
Gender	F	78 (39)			3	144 (28)
	M	120 (61)			4	9 (5)
Family history of CRC	(-)	19 (10)		N	0	104 (53)
	(+)	179 (90)			1	49 (25)
Location†	Distal colon	156 (79)			2	45 (22)
	Proximal colon	43 (22)		M	0	173 (87)
Differentiation	Well/Moderate	167 (84)			1	25 (13)
	Poorly	31 (16)	AJCC stage		I	36 (18)
Microsatellite instability (MSI)	MSS/MSI-L	163 (90)			II	61 (31)
	MSI-H	19 (90)			III	81 (41)
					IV	20 (10)

CRC, colorectal cancer; MSI, microsatellite instability; MSS, microsatellite stable; MSI-L, MSI-low; MSI-H, MSI-high; AJCC, American Joint Committee on Cancer.

† The distal colon was defined as the rectum, sigmoid, and descending colon, whereas the proximal colon was the splenic flexure and more proximal portions of the colon. One patient had double primary cancer in the proximal and distal colon

‡ MSI status was evaluated in 182 patients

RESULTS

Replication Study for the European CRC-susceptibility SNPs in the Korean population

Among the 5 European CRC-susceptibility SNPs, rs4939827, SNP in *SMAD7*, was significantly associated with CRC risk in Korean samples (Table 4). The OR for the additive model was 0.67 (95% CI, 0.47-0.95; p=.024), and the OR for the dominant model was 0.59 (95% CI, 0.39-0.91; p=.016).In the subgroup analysis, the association was more prominent for distal CRC (additive model: OR, 0.65; 95% CI, 0.45-0.96; p=.031, recessive model: OR, 0.62; 95% CI, 0.39-0.98; p=.039), although the additive model showed a significant association for proximal CRC. Interestingly, in the clinical association portion of this study, rs4939827 had a significant association in advanced CRC patients with stage III/IV status, lymph node, or distant metastasis (Table 5).

Among female subjects, rs4779584 was associated with CRC in the additive model (OR, 1.85; 95% CI, 1.00-3.83; p=.051) and in the dominant model (OR, 2.50; 95% CI, 1.16-5.39; p=.020) (Table 4). Rs10795668 was associated with CRC in the additive model (OR, 0.61; 95% CI, 0.39-0.95, p=.029) and recessive model (OR, 0.21; 95% CI, 0.08-0.59, p=.003) in males. In addition, rs10795668 showed a trend of negative association for CRC in the additive model (OR, 0.73; 95% CI, 0.53-1.01; p=.058) and in the dominant model (OR, 0.67; 95% CI, 0.44-1.03; p=.066) across

both genders. However, rs3802842 and rs6983267 did not reveal any significant association between study and control groups.

Genome-wide screening for CRC-susceptibility genetic variants in Koreans

The genome-wide association study was performed using Illumina HumanHap 370k/610k BeadChips among 105 Korean CRC patients. Because the genetic marker-trait associations detected in GWAS may have a larger effect on early-onset traits, (29) we performed genome-wide screening on early-onset CRC patients (the mean age of CRC patients in Illumina HumanHap 370k BeadChips assay and Illumina HumanHap 610k BeadChips assay were 45.5 ± 6.7 years and 38.4 ± 7.2 years, respectively). According to the maximum ratio of the minor allele frequencies (MAF) of CRC in Illumina HumanHap 370k or 610K/MAF of the Korean HapMap, we selected 8 candidate CRC-susceptibility SNPs (rs12266240, rs10491619, rs10941887, rs1859915, rs727235, rs17051076, rs1599695, and rs902960) (Table 6).

Replication Study for the GWAS-identified CRC-susceptibility SNPs

A replication study for GWAS-identified CRC-susceptibility SNPs, using the TaqMan SNP Genotyping Assays, was performed on 189 CRC cases and 190 controls. The CRC samples in this portion of the study was the same as that used in the replication portion of study for European CRC-susceptibility SNPs, but there were 10 CRC cases which could not be genotyped due to an insufficient amount of samples. The clinical characteristics of the CRC group did not differ significantly to Table 3 and are described in detail at Table 7. Because there was a significant differences in

age and gender between both study and control groups (p<.001), the logistic regression analysis was adjusted for these potentially confounding variables.

The additive, dominant, and recessive genetic model of candidate GWAS-identified CRC susceptibility SNPs did not show any significant association with the overall risk of CRC in the population studied (Table 8).

However, the clinical association analysis showed a significant association between rs17051076 and the MSI-H CRC (Table 9). The OR for the additive model was 4.25 (95% CI, 1.51-11.98; p=.006) and the OR of the dominant model was 3.52 (95% CI, 1.13-10.94; p=.030). Furthermore, the additive model of rs17051076 showed a tendency for increased risk of poorly differentiated CRC (OR, 2.45; 95% CI, 0.99-6.09; p=.054). There were no significant association with MSI-L/MSS CRC (additive model: OR, 1.33; 95% CI, 0.71-2.48; p=.369, dominant model: OR, 1.30, 95% CI, 0.70-2.40, p=.410).

Table 4. Replication study for the European CRC-susceptibility SNPs in Korean CRC cases and healthy controls.

SNP	Alleles (1/2)	Controls, n			Cases, n			Additive model†		Dominant model†		Recessive model†	
		11	12	22	11	12	22	OR (95% CI)	*p*	OR (95% CI)	*p*	OR (95% CI)	*p*
rs3802842	A/C	116	154	56	62	94	42	1.13 (0.85-1.51)	.393	1.17 (0.76-1.80)	.478	1.19 (0.71-2.00)	.519
rs4779584	T/C	129	44	7	139	50	4	1.07 (0.70-1.65)	.756	1.18 (0.72-1.95)	.508	0.54 (0.13-2.29)	.400
rs4939827	C/T	182	127	19	126	63	9	**0.67 (0.47-0.95)**	**.024**	**0.59 (0.39-0.91)**	**.016**	0.71 (0.28-1.79)	.463
rs6983267	T/G	101	158	69	59	99	40	1.02 (0.76-1.36)	.901	1.17 (0.75-1.83)	.495	0.87 (0.53-1.43)	.579
rs10795668	G/A	126	155	41	89	86	21	0.73 (0.53-1.01)	.058	0.67 (0.44-1.03)	.066	0.69 (0.35-1.35)	.278
Male													
SNP	**Alleles (1/2)**	**Controls, n**			**Cases, n**			**Additive model‡**		**Dominant model‡**		**Recessive model‡**	
		11	**12**	**22**	**11**	**12**	**22**	**OR (95% CI)**	***p***	**OR (95% CI)**	***p***	**OR (95% CI)**	***p***
rs3802842	A/C	42	64	26	36	57	27	1.08 (0.73-1.60)	.668	1.15 (0.63-2.09)	.648	1.07 (0.54-2.13)	.855
rs4779584	T/C	89	24	3	65	24	4	0.69 (0.38-1.22)	.201	0.66 (0.34-1.29)	.222	0.52 (0.09-3.10)	.516
rs4939827	C/T	73	52	7	75	40	5	0.70 (0.43-1.13)	.140	0.63 (0.36-1.12)	.118	0.73 (0.20-2.76)	.645
rs6983267	T/G	40	56	36	38	58	24	0.89 (0.61-1.29)	.526	0.98 (0.54-1.79)	.950	0.71 (0.37-1.37)	.304
rs10795668	G/A	52	58	21	53	58	8	**0.61(0.39-0.95)**	**.029**	0.75 (0.43-1.33)	.330	0.21 (0.08-0.59)	.003
Female													
SNP	**Alleles (1/2)**	**Controls, n**			**Cases, n**			**Additive model‡**		**Dominant model‡**		**Recessive model‡**	
		11	**12**	**22**	**11**	**12**	**22**	**OR (95% CI)**	***p***	**OR (95% CI)**	***p***	**OR (95% CI)**	***p***
rs3802842	A/C	74	90	30	26	37	15	1.18 (0.77-1.82)	.441	1.18 (0.63-2.22)	.609	1.37 (0.63-3.01)	.430
rs4779584	T/C	50	26	1	64	20	3	1.85 (1.00-3.83)	.051	**2.50 (1.16-5.39)**	**.020**	0.58 (0.05-6.72)	.664
rs4939827	C/T	109	75	12	51	23	4	0.63 (0.38-1.06)	.083	0.55 (0.29-1.03)	.063	0.67 (0.18-2.47)	.552
rs6983267	T/G	61	102	33	21	41	16	1.22 (0.79-1.90)	.375	1.46 (0.73-2.89)	.281	1.12 (0.52-2.42)	.766
rs10795668	G/A	74	97	20	36	28	13	0.91 (0.57-1.45)	.687	0.59 (0.32-1.11)	.100	2.15 (0.88-5.28)	.094

CRC, colorectal cancer; Chr, chromosome; SNP, single-nucleotide polymorphism; OR, odds ratio; CI, confidence intervals.

† Adjusted for age and gender

‡ Adjusted for age

Bold character indicates statistical significance P<0.05

Table 5. Clinical association analysis for the European CRC-susceptibility SNPs in Korean CRC cases and healthy controls.

SNP	Young onset CRC (diagnosed <50 years of age) vs. Controls								MSI- H CRC vs. Controls							
	Young-onset CRC, n			Additive model†		Dominant model†		Recessive model†		MSI-H CRC, n			Additive model†		Dominant model†		Recessive model†	
	11	12	22	OR (95% CI)	*p*	OR (95% CI)	*p*	OR (95% CI)	*p*	11	12	22	OR (95% CI)	*p*	OR (95% CI)	*p*	OR (95% CI)	*P*
rs3802842	11	16	7	1.05 (0.62-1.78)	.870	0.92 (0.44-2.11)	.916	1.22 (0.49-3.04)	.675	6	8	5	1.26 (0.63-2.49)	.513	1.32 (0.44-3.95)	.625	1.42 (0.45-4.49)	.549
r4779584	21	9	1	0.86 (0.40-1.83)	.858	0.94 (0.37-2.39)	.942	.042 (0.04-4.12)	.423	11	8	0	1.55 (0.63-3.83)	.341	1.98 (0.69-5.64)	.204	-	-
rs4939827	20	13	1	0.94 (0.50-1.76)	.837	1.01 (0.48-2.14)	.973	0.52 (0.06-4.19)	.539	11	7	1	0.71 (0.03-1.69)	.443	0.67 (0.24-1.83)	.420	0.73 (0.08-6.43)	.778
rs6983267	11	15	8	0.97 (0.58-1.63)	.921	0.87 (0.40-1.92)	.733	1.10 (0.46-2.63)	.837	7	10	2	0.91 (0.46-1.82)	.795	1.26 (0.41-3.88)	.689	0.54 (0.14-2.11)	.380
rs10795668	16	15	2	0.77 (0.44-1.34)	.351	0.86 (0.40-1.83)	.860	0.37 (0.08-1.67)	.197	7	10	2	0.95 (0.44-2.04)	.894	1.00 (0.35-2.84)	1.000	0.80 (0.16-4.05)	.791

SNP	Proximal CRC vs. Controls								Distal CRC vs. Controls							
	Proximal CRC, n			Additive model†		Dominant model†		Recessive model†		Distal CRC, n			Additive model†		Dominant model†		Recessive model†	
	11	12	22	OR (95% CI)	*p*	OR (95% CI)	*p*	OR (95% CI)	*p*	11	12	22	OR (95% CI)	*p*	OR (95% CI)	*p*	OR (95% CI)	*P*
rs3802842	13	21	9	1.03 (0.63-1.68)	.912	1.05 (0.50-2.21)	.891	1.02 (0.42-2.48)	.973	49	74	33	1.16 (0.85-1.58)	.366	1.20 (0.75-1.93)	.443	1.23 (0.70-2.16)	.477
r4779584	30	11	1	0.83 (0.41-1.70)	.618	0.88 (0.39-1.98)	.748	0.42 (0.03-5.30)	.499	109	40	3	1.20 (0.76-1.90)	.433	1.35 (0.79-2.31)	.267	0.85 (0.14-2.98)	.575
rs4939827	29	11	3	0.61 (0.33-1.14)	.612	**0.46 (0.22-0.98)**	**.045**	1.18 (0.30-4.66)	.809	98	52	6	**0.65 (0.45-0.96)**	**.031**	**0.62 (0.39-0.98)**	**.039**	0.52 (0.18-1.53)	.236
rs6983267	11	25	7	0.93 (0.57-1.50)	.755	1.25 (0.57-2.75)	.572	0.58 (0.23-1.47)	.251	48	75	33	1.05 (0.77-1.42)	.780	1.19 (0.74-1.94)	.472	0.92 (0.53-1.57)	.750
rs10795668	17	22	3	0.72 (0.42-1.25)	.247	0.79 (0.38-1.63)	.525	0.40 (0.11-1.50)	.176	72	65	18	0.75 (0.53-1.06)	.101	0.66 (0.42-1.03)	.069	0.80 (0.40-1.63)	.542

SNP	Poorly differentiated CRC vs. Controls								AJCC stage III or IV CRC vs. Controls							
	Poorly differentiated CRC, n			Additive model†		Dominant model†		Recessive model†		AJCC stage III or IV, n			AJCC stage III or IV, n		AJCC stage III or IV, n		Recessive model†	
	11	12	22	OR (95% CI)	11	11	11	OR (95% CI)	*p*	11	12	22	OR (95% CI)	*p*	OR (95% CI)	*p*	OR (95% CI)	*P*
rs3802842	9	15	7	1.22 (0.70-2.13)	.484	1.34 (0.57-3.17)	.507	1.26 (0.47-3.37)	.641	50	79	33	1.12 (0.82-1.53)	.478	1.21 (0.76-1.93)	.421	1.09 (0.62-1.92)	.758
r4779584	23	8	0	0.85 (0.36-2.01)	.854	0.84 (0.37-2.39)	.890	-	-	114	41	4	1.12 (0.71-1.77)	.613	1.23 (0.73-2.09)	.433	0.67 (0.15-2.92)	.591
rs4939827	21	9	1	0.48 (0.27-1.03)	.058	0.43 (0.18-1.02)	.055	0.40 (0.05-3.42)	.401	103	52	7	**0.63 (0.43-0.93)**	**.020**	**0.57 (0.36-0.90)**	**.017**	0.61 (0.22-1.68)	.335

rs6983267	8	17	6	1.03 (0.59-1.79)	.914	1.43 (0.57-3.58)	.447	0.72 (0.26-1.97)	.517	46	85	31	1.00 (0.73-1.36)	.976	1.23 (0.75-2.00)	.410	0.76 (0.44-1.32)	.333
rs10795668	15	13	3	0.63 (0.33-1.21)	.163	0.52 (0.23-1.17)	.112	0.74 (0.19-2.85)	.665	70	69	21	0.80 (0.57-1.12)	.184	0.70 (0.44-1.10)	.124	0.87 (0.44-1.72)	.688

SNP	**CRC with lymph node metastasis vs. Controls**									**CRC with distant metastasis vs. Controls**							
	Lymph node metastasis , n			Additive model†		Dominant model†		Recessive model†		Distant metastasis, n			Additive model†		Dominant model†		Recessive model†	
	11	12	22	OR (95% CI)	*p*	OR (95% CI)	*p*	OR (95% CI)	*p*	11	12	22	OR (95% CI)	p	OR (95% CI)	p	OR (95% CI)	P
rs3802842	31	50	13	0.90 (0.61-1.32)	.585	1.02 (0.59-1.77)	.956	0.65 (0.30-1.40)	.650	6	16	3	1.01 (0.54-1.88)	.986	1.48 (0.56-3.94)	.434	0.54 (0.14-2.02)	.357
r4779584	68	22	2	1.11 (0.64-1.91)	.721	1.19 (0.63-2.23)	.597	0.75 (0.12-4.62)	.754	20	4	0	0.56 (0.19-1.68)	.303	0.58 (0.18-1.88)	.365	-	-
rs4939827	62	30	2	**0.50 (0.61-0.82)**	**.006**	**0.48 (0.27-0.84)**	**.011**	0.26 (0.05-1.27)	.095	21	4	0	**0.22 (0.08-0.66)**	**.007**	**0.21 (0.07-0.64)**	**.007**	-	-
rs6983267	28	46	20	0.99 (0.69-1.42)	.954	1.14 (0.64-2.03)	.658	0.83 (0.44-1.57)	.558	4	16	5	1.23 (0.68-2.22)	.488	2.52 (0.81-7.87)	.112	0.73 (0.25-2.15)	.568
rs10795668	39	40	14	0.85 (0.57-1.27)	.427	0.74 (0.43-1.27)	.274	1.00 (0.46-2.18)	.997	7	12	6	1.42 (0.76-2.68)	.273	1.39 (0.54-3.59)	.497	1.87 (0.65-5.38)	.247

CRC, colorectal cancer; SNP, single-nucleotide polymorphism; MAF, minor allele frequency; OR, odds ratio; CI, confidence intervals; MSI-H, microsatellite instability-high; AJCC , American Joint Committee on Cancer.

†Adjusted for age and gender

Bold character indicates statistical significance P<0.05

Table 6. CRC-susceptibility SNP selection using Illumina HumanHap 370K and Illumina HumanHap 610K beadchip arrays.

	MAF in Koran HapMap	MAF in Korean CRC (Illumina HumanHap 370K, n=45)	MAF in Korean CRC (Illumina HumanHap 610K, n=60)	MAF of CRC in 370K / MAF of Korean HapMap	MAF of CRC in 610K / MAF of Korean HapMap
rs12266240	0.010	0.191	0.125	19.100	12.500
rs10491619	0.028	0.090	0.075	3.214	2.679
rs10941887	0.050	0.145	0.125	2.900	2.500
rs1859915	0.072	0.208	0.150	2.889	2.083
rs727235	0.035	0.100	0.100	2.857	2.857
rs17051076	0.052	0.138	0.125	2.654	2.404
rs1599695	0.076	0.159	0.167	2.092	2.197
rs902960	0.033	0.068	0.075	2.061	2.273

CRC, colorectal cancer; SNP, single-nucleotide polymorphism; MAF, minor allele frequency.

Table 7. Demographic and clinicopathological characteristics of 189 colorectal cancer patients in GWAS-identified candidate CRC-susceptibility SNPs.

Variables		N	%	Variables			n	%
Age at diagnosis	≥ 50 y	158	84	TNM classification	T	1	16	9
	< 50 y	31	16			2	25	13
Gender	F	74	39			3	140	74
	M	115	61			4	8	4
Family history of CRC	(+)	19	10		N	0	101	53
	(-)	170	90			1	47	25
Location†	Distal CRC	149	79			2	41	22
	Proximal CRC	40	21		M	0	166	88
Microsatellite instability	MSS/MSI-L	156	85			1	23	12
(MSI)‡	MSI-H	19	15	AJCC stage	I		34	18
Differentiation	Well/Moderate	161	85		II		60	32
	Poorly	28	15		III		76	40
					IV		19	10

GWAS, genome-wide association study; CRC, colorectal cancer; MSI, microsatellite instability; MSS, microsatellite stable; MSI-L, MSI-low; MSI-H, MSI-high; AJCC , American Joint Committee on Cancer.

† The distal colon was defined as the rectum, sigmoid, and descending colon, whereas the proximal colon was the splenic flexure and more proximal portions of the colon. One patient had double primary cancer in the proximal and distal colon

‡ MSI status was evaluated in 175 patients

Table 8. Replication study for the GWAS-identified CRC susceptibility variants in Korean CRC cases and healthy controls.

SNP	Alleles	Controls, n			Cases, n			Additive model†		Dominant model†		Recessive model†	
	(1/2)	11	12	22	11	12	22	OR (95% CI)	*p*	OR (95% CI)	*p*	OR (95% CI)	*p*
rs902960	A/G	175	14	0	170	15	2	1.21 (0.56-2.61)	.627	1.10 (0.47-2.55)	.831	-	-
rs1599695	T/C	147	42	1	152	35	2	0.86 (0.50-1.47)	.580	0.81 (0.46-1.45)	.483	1.77 (0.14-21.84)	.656
rs1859915	T/C	135	51	4	132	52	4	0.82 (0.51-1.31)	.402	0.83 (0.49-1.40)	.476	0.55 (0.10-2.93)	.482
rs727235	G/A	165	16	1	161	17	0	0.78 (0.35-1.75)	.546	0.88 (0.48-2.06)	.775	-	-
rs10491619	A/G	153	33	4	149	36	0	0.85 (0.49-1.46)	.557	0.92 (0.51-1.65)	.777	-	-
rs10941887	A/G	139	47	2	144	40	3	0.92 (0.55-1.54)	.916	0.88 (0.50-1.53)	.649	1.63 (0.16-16.46)	.679
rs12266240	T/C	145	41	2	141	43	4	1.34 (0.82-2.33)	.232	1.46 (0.83-2.57)	.190	0.96 (0.12-7.87)	.968
rs17051076	G/A	154	35	0	140	42	2	1.45 (0.81-2.57)	.210	1.37 (0.77-2.47)	.288	-	-
						Male							
SNP	Alleles	Controls, n			Cases, n			Additive model‡		Dominant model‡		Recessive model‡	
	(1/2)	11	12	22	11	12	22	OR (95% CI)	*p*	OR (95% CI)	*p*	OR (95% CI)	*p*
rs902960	A/G	57	6	0	102	11	1	1.40 (0.48-3.82)	.561	1.24 (0.41-3.82)	.704	-	-
rs1599695	T/C	45	17	1	92	21	2	0.80 (0.39-1.61)	.524	0.70 (0.32-1.56)	.386	1.73 (0.14-21.11)	.670
rs1859915	T/C	44	18	1	81	33	1	0.70 (0.34-1.45)	.340	0.77 (0.36-1.64)	.494	0.10 (0.01-1.93	.129
rs727235	G/A	51	7	1	94	13	0	0.93 (0.33-2.61)	.930	1.17 (0.38-3.64)	.781	-	
rs10491619	A/G	50	10	3	25	3	0	0.38 (0.11-1.30)	.122	0.37 (0.09-1.48)	.161	-	
rs10941887	A/G	46	16	0	91	23	1	0.91 (0.41-2.04)	.820	0.90 (0.40-2.03)	.800	-	-
rs12266240	T/C	50	12	1	95	29	1	1.37 (0.65-2.89)	.405	1.48 (0.64-3.45)	.360	0.89 (0.18-12.11)	.887
rs17051076	G/A	50	13	0	87	24	2	1.30 (0.59-2.84)	.518	1.21 (0.68-2.13)	.515	-	

Female													
SNP	Alleles	Controls, n			Cases, n			Additive model‡		Dominant model‡		Recessive model‡	
	(1/2)	11	12	22	11	12	22	OR (95% CI)	*p*	OR (95% CI)	*P*	OR (95% CI)	*p*
rs902960	A/G	118	8	0	68	4	1	1.02 (0.32-3.27)	.980	1.11 (0.24-3.36)	.904	-	-
rs1599695	T/C	102	25	0	60	14	0	0.84 (0.41-2.17)	.888	0.94 (0.41-2.17)	.888	-	-
rs1859915	T/C	91	33	3	51	19	3	0.91 (0.49-1.71)	.769	0.88 (0.42-1.85)	.743	0.96 (0.15-6.27)	.965
rs727235	G/A	114	9	0	67	4	0	0.78 (0.18-3.40)	.744	0.78 (0.18-3.40)	.744	-	-
rs10491619	A/G	103	23	1	124	33	0	0.94 (0.47-1.88)	.860	0.99 (0.48-2.03)	.980	-	-
rs10941887	A/G	93	31	2	53	17	2	0.90 (0.45-1.79)	.766	0.84 (0.39-1.82)	.839	1.56 (0.14-17.28)	.719
rs12266240	T/C	89	22	4	52	21	0	1.39 (0.66-2.93)	.389	1.44 (0.67-3.09)	.356	-	-
rs17051076	G/A	103	22	0	53	18	0	1.63 (0.71-3.75)	.254	1.63 (0.71-3.75)	.254	-	-

CRC, colorectal cancer; Chr, chromosome; SNP, single-nucleotide polymorphism; OR, odds ratio; CI, confidence intervals.

† Adjusted for age and gender

‡ Adjusted for age

Bold character indicates statistical significance P<0.05

Table 9. Clinical association analysis for the GWAS-identified CRC-susceptibility SNPs in Korean CRC cases and healthy controls.

SNP	Young onset CRC (diagnosed <50 years of age) vs. Controls									MSI-H CRC vs. Controls								
	Young-onset CRC, n			Additive model†		Dominant model†		Recessive model†		MSI-H CRC, n			Additive model†		Dominant model†		Recessive model†	
	11	12	22	OR (95% CI)	*p*	OR (95% CI)	*p*	OR (95% CI)	*p*	11	12	22	OR (95% CI)	*p*	OR (95% CI)	*p*	OR (95% CI)	*p*
rs902960	28	3	0	1.58 (0.39-6.43)	.520	1.58 (0.39-6.43)	.520	-	-	17	2	0	1.30 (0.25-6.81)	.753	1.30 (0.25-6.81)	.753	-	-
rs1599695	27	3	1	0.57 (0.20-1.57)	.275	0.42 (0.13-1.34)	.142	4.26 (0.25-72.80)	.317	15	4	0	0.90 (0.26-3.07)	.866	0.92 (0.26-3.21)	.892	-	-
rs1859915	22	8	1	1.37 (0.61-3.06)	.449	1.25 (0.51-3.01)	.628	4.99 (0.40-62.12)	.212	14	4	1	0.86 (0.33-2.23)	.749	0.77 (0.24-2.49)	.658	1.13 (0.11-12.16)	.917
rs727235	26	4	0	0.96 (0.26-3.64)	.955	0.98 (0.25-3.81)	.975	-	-	17	1	0	0.52 (0.07-3.90)	.527	0.54 (0.06-4.79)	.577	-	-
rs10491619	25	6	0	0.85 (0.35-2.10)	.732	1.01 (0.36-2.80)	.987	-	-	17	2	0	0.52 (0.11-2.45)	.411	0.35 (0.05-2.72)	.316	0.34 (0.04-3.01)	.332
rs10941887	27	4	0	0.35 (0.11-1.08)	.067	0.34 (0.11-1.09)	.069	-	-	16	2	1	0.65 (0.20-2.07)	.461	0.47 (0.11-1.96)	.303	3.56 (0.08-159.25)	.513
rs12266240	22	9	0	1.10 (0.46-2.60)	.837	1.18 (0.47-2.94)	.726	-	-	12	5	1	2.14 (0.77-5.98)	.146	2.55 (0.79-8.26)	.119	2.02 (0.08-51.69)	.671
rs17051076	24	6	0	0.89 (0.31-2.51)	.824	0.89 (0.31-2.51)	.824	-	-	10	8	1	4.25 (1.51-11.98)	.006	3.52 (1.13-10.94)	.030	-	-

SNP	Proximal CRC vs. Controls									Distal CRC vs. Controls								
	Proximal CRC, n			Additive model†		Dominant model†		Recessive model†		Distal CRC, n			Additive model†		Dominant model†		Recessive model†	
	11	12	22	OR (95% CI)	*p*	OR (95% CI)	*p*	OR (95% CI)	*p*	11	12	22	OR (95% CI)	*p*	OR (95% CI)	*p*	OR (95% CI)	*p*
rs902960	35	5	0	2.01 (0.61-6.67)	.254	2.01 (0.61-6.67)	.254	-	-	135	10	2	0.87 (0.43-2.20)	.939	0.79 (0.31-2.01)	.616	-	-
rs1599695	28	12	0	1.30 (0.57-2.99)	.538	1.36 (0.58-3.22)	.483	-	-	124	23	2	0.78 (0.43-1.40)	.398	0.70 (0.37-1.33)	.276	2.33 (0.19-29.00)	.510
rs1859915	26	13	1	1.02 (0.50-2.09)	.953	1.13 (0.50-2.59)	.766	0.50 (0.05-5.38)	.567	106	39	3	0.14 (0.40-1.14)	.677	0.67 (0.37-1.19)	.170	0.45 (0.07-2.86)	.403
rs727235	34	2	0	0.41 (0.09-1.92)	.261	0.42 (0.08-2.20)	.301	-	-	127	15	0	0.87 (0.37-2.03)	.746	1.00 (0.41-2.44)	.995	-	-
rs10491619	35	4	0	0.47 (0.15-1.52)	.208	0.47 (0.15-1.52)	.208	-	-	114	32	0	0.96 (0.55-1.69)	.888	1.05 (0.57-1.93)	.881	-	-
rs10941887	31	6	2	0.90	.815	0.74	.550	4.83	.282	113	34	1	0.98	.946	0.97	.906	1.37	.823

				(0.39-2.11)		(0.28-1.97)		(0.28-84.91)					(0.56-1.72)		(0.54-1.74)		(0.09-21.54)	
rs12266240	31	7	1	1.22	.657	1.32	.575	0.69	.831	110	36	3	1.35	.283	1.41	.256	1.18	.883
				(0.51-2.92)		(0.50-3.46)		(0.02-20.22)					(0.78-2.34)		(0.78-2.56)		(0.14-10.04)	
rs17051076	27	12	0	2.16	.094	2.16	.094	-	-	113	30	2	1.22	.524	1.18	.437	-	-
				(0.88-5.30)		(0.88-5.30)							(0.66-2.25)		(0.78-1.81)			

SNP	Poorly differentiated CRC vs. Controls									AJCC stage III or IV CRC vs. Controls								
	Poorly differentiated CRC, n			Additive model†		Dominant model†		Recessive model†		AJCC stage III or IV, n			AJCC stage III or IV, n		AJCC stage III or IV, n		Recessive model†	
	11	12	22	OR (95% CI)	11	11	11	OR (95% CI)	*p*	11	12	22	OR (95% CI)	*P*	OR (95% CI)	*p*	OR (95% CI)	*p*
rs902960	24	4	0	1.54	.534	1.54	.534	-	-	139	13	2	1.28	.544	1.15	.765	-	-
				(0.40-5.93)		(0.40-5.93)							(0.58-2.85)		(0.47-2.78)			
rs1599695	21	7	0	1.09	.864	1.12	.828	-	-	121	32	2	1.02	.957	0.97	.912	2.34	.509
				(0.40-2.97)		(0.40-3.13)							(0.58-1.78)		(0.53-1.77)		(0.19-29.31)	
rs1859915	21	5	2	0.68	.360	0.50	.196	1.33	.765	105	46	3	0.80	.372	0.85	.572	0.61	.183
				(0.30-1.55)		(0.78-1.43)		(0.20-8.69)					(0.48-1.32)		(0.49-1.49)		(0.06-1.74)	
rs727235	26	1	0	0.35	.316	0.34	.332	-	-	133	12	0	0.66	.351	0.74	.527	-	-
				(0.05-2.72)		(0.04-3.01)							(0.28-1.58)		(0.29-1.88)			
rs10491619	24	4	0	0.57	.362	0.58	.388	-	-	119	33	0	0.94	.827	1.02	.939	-	-
				(0.17-1.89)		(0.17-1.99)							(0.53-1.66)		(0.56-1.89)			
rs10941887	19	8	1	1.44	.418	1.35	.539	4.37	.347	118	33	3	091	.743	0.86	.624	1.93	.586
				(0.60-3.44)		(0.52-3.54)		(0.20-94.80)					(0.53-1.58)		(0.48-1.56)		(0.18-20.47)	
rs12266240	24	3	0	0.60	.430	0.61	.460	-	-	114	36	4	1.48	.166	1.57	.139	1.20	.868
				(0.17-2.14)		(0.16-2.27)							(0.85-2.58)		(0.86-2.87)		(0.14-10.04)	
rs17051076	18	8	1	2.45	.054	1.97	.188	-	-	116	35	1	1.42	.262	1.63	.327	-	-
				(0.99-6.09)		(0.72-5.40)							(0.77-2.63)		(0.73-2.54)			

SNP	CRC with lymph node metastasis vs. Controls									CRC with distant metastasis vs. Controls								
	Lymph node metastasis , n			Additive model†		Dominant model†		Recessive model†		Distant metastasis, n			Additive model†		Dominant model†		Recessive model†	
	11	12	22	OR (95% CI)	*p*	OR (95% CI)	*p*	OR (95% CI)	*p*	11	12	22	OR (95% CI)	p	OR (95% CI)	p	OR (95% CI)	p
rs902960	81	5	1	0.84	.733	0.70	.545	-	-	19	3	1	2.61	.096	2.20	.238	-	-
				(0.31-2.60)		(0.27-2.19)							(0.84-8.05)		(0.60-8.09)			
rs1599695	33	21	1	1.08	.828	1.01	.969	3.31	.406	15	8	0	1.76	.251	1.88	.219	-	-
				(0.55-2.11)		(0.50-2.07)		(0.20-55.55)					(0.67-4.64)		(0.69-5.17)			
rs1859915	60	25	2	0.74	.320	0.80	.501	0.27	.197	14	9	0	0.80	.635	0.96	.942	-	-
				(0.41-1.34)		(0.41-1.55)		(0.04-1.98)					(0.32-2.00)		(0.35-2.66)			
rs727235	72	9	0	0.84	.718	0.98	.972	-	-	16	5	0	1.52	.480	1.95	.306	-	-
				(0.31-2.22)		(0.34-2.82)							(0.47-4.91)		(0.54-7.03)			
rs10491619	67	19	0	1.03	.930	1.13	.749		-	15	8	0	1.73	.247	1.99	.184	-	-

				(0.53-2.01)		(0.55-2.31)							(0.69-4.35)		(0.72-5.48)			
rs10941887	70	17	0	0.79	.529	0.80	.555	-	-	21	2	0	0.29	.118	0.28	.118	-	-
				(0.38-1.64)		(0.38-1.67)							(0.06-1.38)		(0.06-1.38)			
rs12266240	67	19	2	1.39	.320	1.43	.336	1.80	.613	18	5	0	1.14	.806	1.24	.708	-	-
				(0.73-2.65)		(0.69-2.94)		(0.18-17.70)					(0.39-3.33)		(0.40-3.89)			
rs17051076	68	17	1	0.99	.986	1.12	.473	-	-	17	5	1	1.99	.181	1.47	.478	-	-
				(0.48-2.08)		(0.82-1.55)							(0.73-5.42)		(0.51-4.27)			

GWAS, genome-wide association study; CRC, colorectal cancer; SNP, single-nucleotide polymorphism; MAF, minor allele frequency; OR, odds ratio; CI, confidence intervals; MSI-high, microsatellite instability-high; AJCC , American Joint Committee on Cancer.

[†]Adjusted for age and gender

Bold character indicates statistical significance P<0.05

DISCUSSION

In the past few decades, the incidence of CRC has been increasing rapidly in developed Asian countries. As such, studies on the European GWAS-identified CRC-susceptibility SNPs has been replicated in the respective populations of Japan (30, 31), Singapore (32), Hong Kong (33) and China (34). Although the incidence of CRC in Korea is reported to be as high as that of developed Western countries (http://globocan.iarc.fr), there has been no study to validate the use of European GWAS-identified CRC-susceptibility SNPs or genome-wide screening for CRC susceptibility SNPs in the Korean population.

In the present study, we have evaluated and confirmed the association of European GWAS-identified CRC-susceptibility SNPs, rs4939827, rs4779584, and rs10795668, with CRC risk in the Korean population. In addition, using genome-wide screening, candidate CRC-susceptibility SNPs were selected in Korean CRC patients, and the novel rs17051076 was found to be associated with MSI-H CRC.

The rs4939827 is located in the intron region of the *SMAD7* gene on 18q21. A large-scaled study and a meta-analysis (16, 19) revealed that the OR showed a decreased risk for the minor rs4939827 allele with highly statistical significance (16), which were similar to the results of present study. The association of rs4939827 with the risk of CRC had already been replicated in a number of Asian populations (30, 34, 35). Since rs4939827 was identified as a CRC-susceptibility SNP in the GWAS (16),

the biological rationale for the association between SMAD7 gene and CRC had been explored. The SMAD7 is involved in inflammation-related pathways and has been shown to modulate TGF-β and Wnt signaling (36), which are central to the development of CRCs (37). It has been shown that the over-expression of SMAD7 can induce tumorigenesis by blocking TGF-β-induced growth inhibition and apoptosis, which suggests a possible linkage to the association between rs4939827 SNP and CRC (38). In addition, previous studies have reported that site-specific differences of rs4939827, with associations observed in distal colon tumors but not in proximal tumors (37, 39). In the present study, this SNP also showed a prominent association to distal CRC in contrast to proximal CRC.

Additionally, the rs4779584 and rs10795668 showed notable gender-specific differences; rs4779584 showed a significant association only in females and rs10795668 showed a significant association only in males. Rs4779584 is located in the CRAC1 locus on chromosome 15q14 (18), and the association between this SNP to CRC had been replicated in Hong Kong Chinese (33). Although the risk allele in our population is the same as that of the European population, T is a major allele (0.85) in our control population, which is a similar MAF to that in Hong Kong Chinese study (33). A previous meta-analysis and replication study showed a strong association of rs4779584 with CRC risk in both genders (18, 33), whereas in this study, a significant association was found only among the male. On the other hand, the rs10795668, located on chromosome 10p14, showed a significant association with

CRC risk only in the female. The rs10795668 showed the strongest association with CRC risk in the European (21) and Chinese populations (33, 34, 40); however, there are no known protein-coding transcripts associated with rs10795668 (40).

In contrast to rs4939827, rs4779584, and rs10795668, we did not observe any association between CRC risk and the remaining two European CRC-susceptibility SNPs, rs6983267 and rs3802842. Previous studies have reported that rs6983267 has been associated with numerous cancers including CRC (41), and suggested a biomolecular interaction between the rs6983267 region and MYC, an important proto-oncogene that is over-expressed in CRC (42). Rs3802842 is a SNP in the C11orf93 gene on chromosome 11q23.1. The rs6983267 and rs3802842 SNPs are replicated in other Asian populations, such as those in Japanese (30) and Chinese populations (33-35). The incongruent results between this study and that of previous meta-analysis of GWAS could be attributed to the racial and ethnic differences of the respective study populations because allele frequencies of these SNPs may very well be different. Another possible reason that these SNPs were not found to be susceptibility variants for CRC might be that the sample size was large enough to provide of sufficient statistical power in this study.

After performing the replication study on European CRC-susceptibility SNPs, we performed genome-wide screening for CRC patients and, subsequently, 8 candidate SNPs were selected upon comparing the screening results to the Korean HapMap

database. The rs1859915 is located in the protein coding SLC8A3 gene. Two SNPs, rs12266240 and rs727235, are located in the transcript processed genes, RP11-494M8.4 and RP11-659O3.1, respectively. The remaining 5 SNPs, rs902960, rs1599695, rs17051076, rs10491619 and rs10941887 are intragenic SNPs which are located at least 5 kb up- or down-stream of a known gene locus (43). Replication analysis of these SNPs identified that rs17051076 is associated with MSI-H CRC, but not with MSI-L or MSS CRC, in the Korean population.

Previously, rs17051076 has not been reported to have any association with MSI-H CRC. The RP11-401I19 gene, a long intergenic non-coding RNA, is located 151Kb downstream from rs17051076, and GAPDHP56, a known pseudogene, is located 225Kb upstream of rs17051076. However, these genes exist faraway from rs17051076 and additional fine-mapping and functional analyses are needed to further address the role this SNP might have in increasing the risk of MSI-H CRC. For MSI-H CRCs, studies on genetic variants had been focused on the polymorphisms and methylation of the mismatch repair gene, MLH1, and its promoter region (44). MSI-H CRCs have a tendency to develop in the proximal colon, with increased tumor-infiltrating lymphocytes, and with a poorly differentiated, mucinous or signet ring histology (45). However, proximal CRC and poorly differentiated histology were not associated rs17051076. It may be explained by indirect relationship between rs17051076 and MSI-H CRC associated clinical features, although direct association can be proposed between rs17051076 and MSI-

H CRC or between MSI-H CRC and clinical features.

In conclusion, several CRC-susceptibility SNPs from European studies were shown to be also associated with increased CRC risk in the Korean population despite the genetic variation across the racial and ethnic diversity. The European CRC-susceptibility SNP, rs4939827, in SMAD7 may also contribute to the risk of CRC in Koreans. In addition, European CRC-susceptibility SNP rs4779584 may be associated with CRC risk in Korean males, and the rs10795668 SNP may be associated with increased risk in Korean females. Despite the small sample size of this replication study, the GWAS-identified novel SNP, rs17051076, was associated with an increased risk of MSI-H CRC in the Korean population. The limitation of our study is that our sample size may not have been large enough to detect minimal associations and interactions. Therefore, the associations between these SNPs and CRC risk remain to be clarified. The results of this study should be approached cautiously, and need validation in future large-scale genome-wide association analyses and replicative experiments.

CHAPTER 2

A common copy number variation in *PKD1L2* is associated with colorectal cancer predisposition

Introduction

In recent years, genome-wide association studies (GWAS) have identified over 20 low-penetrance CRC susceptibility loci (9, 26, 46-48). However, these variants account for only a small fraction of heritability in CRC (49). Furthermore, most GWAS are biased towards the detection of common single nucleotide polymorphisms (SNPs), thereby potentially missing other candidates such as rare SNPs and structural changes (47).

Copy number variations (CNVs) are structural variants in the human genome involving changes in copy number in a particular genomic region (due to deletion or amplification of DNA segments), ranging from a kilobase to several megabases in length (50). The widespread presence of CNVs in the human genome plays an important role in human traits and disorders (51). However, the role of CNVs in disease has been under-investigated because the human genome project did not include the distribution of CNVs across the genome (52, 53). Recently, studies of CNVs have been highlighted as accounting for the missing heritability of complex diseases, and several studies have reported that CNVs are associated with polygenic diseases such as neuroblastoma, systemic lupus erythematosus, schizophrenia, and autism (50, 54-57).

Unfortunately, although the estimated lifetime risk of CRC is as high as approximately 5% in the general population (58), CRC susceptibility CNVs have not yet been systematically evaluated. Several rare CNVs (present at frequencies <1%) such as *APC*, *BMPR1A*, *BUB1*, *BUB3*, *CDH18*, *EPCAM*, *GREM1*, *MAFH4*, *MSH2*,

MSH6, *PMS2*, and *SMAD4* were reported in familial or early-onset CRC (59, 60) (61). However, only a few common CNVs that predispose to CRC have been reported. These include CNVs in 11q11 and the phase II detoxification genes in a European population (62, 63). In the present study, we used array comparative genomic hybridization (aCGH) to perform a genome-wide screen for large CNVs, in order to identify the novel common CNVs that may contribute to the development of sporadic CRC.

Material and Methods

Study population

All participants, both case and control, were of Korean descent. All of the case samples were obtained from the Samsung Medical Center. All of the healthy control samples who were not diagnosed with CRC were obtained from the Sihwa-Banwol Environmental Health Cohort (SBEHC) of Seoul National University (27), and from health check-up population of Seoul National University Hospital, Seoul, Korea. All samples underwent health interviews and medical health examinations. Genomic DNA was extracted from peripheral blood using QuickGene-610L according to the manufacturer's protocol (Fujifilm, Tokyo, Japan). Genomic DNA samples obtained from 1,874 CRC patients and 2,088 normal controls from the Sihwa-Banwol cohort and a health check-up population were collected (Table 10). 36 case samples and 47 controls from the Sihwa-Banwol cohort were used for aCGH experiment. All samples were used for validation and replication of CNV identified in the aCGH experiment (Table 11). Written informed consent was obtained from all study participants. This study was approved by the Institutional Review Boards of Samsung Medical Center, Seoul National University Hospital, and Seoul National University.

Array comparative genomic hybridization

36 CRC and 47 control genomic DNAs were selected from our study subjects for the discovery aCGH experiment (Table 11). SurePrint G3 Custom CGH Microarrays, 4 X 180k which were customized to identify Asian specific CNVs (Agilent Technologies, California, USA) were used for this study (64). All CRC samples, except those from two patients, were from early-onset patients under the age of 50 (mean age: 39.7 ± 7.47). NA10851, one of the HapMap Caucasian samples, was used as a reference in the aCGH experiment. aCGH was performed using a modified Agilent protocol which uses restriction enzymatic methods. The aCGH was performed according to the manufacturer's instructions. The array chips were scanned using an Agilent G2565CA microarray scanner (Agilent Technologies) and analyzed using FEATURE EXTRACTION software 10.5.1.1 (Agilent Technologies).

Array CGH data analysis

Using the CARA software with default parameters, enabling the detection of reference-unbiased absolute CNVs (65), we converted the raw array CGH data of FEATURE EXTRACTION software and used Agilent Genomic Workbench Standard Edition 5.0.14 (Agilent Technologies) to call CNVs. Aberration algorithm ADM-2, centralization threshold 6, bin size 10, and minimum number of probes 5 were used as the analysis settings for CNV calling. The threshold of absolute average log2 ratio

was set at 0.4 and 0.3 for CRC cases and controls, respectively. The human genome assembly NCBI36/hg18 was used as a reference. CNVs were described as segments defined by the mutual overlap of CNV regions (CNVR) from data of Agilent genomic Workbench Standard Edition. CNVs were excluded when the sizes of the deletions or amplifications were < 1 kb (66). For those CNVs found only in CRCs, the number of samples should be at least 3 to be selected as candidates for further experiments. If deletions or amplifications were observed in both sample groups, CNV regions were selected when the odds ratio (OR) between the case and control was over 3 (Table 12).

CNV validation using breakpoint PCR

Breakpoint PCR was performed to determine the exact deletion region in the *PKD1L2* gene. All primers were designed using PRIMER3 software (67). Primers for *PKD1L2* were 5'-GATACTCAGGAGACAATGGAGTCA-3' (forward), 5'-CTAAGACAGGTGAGCGGTAG-3' (reverse for normal type), and 5'-GCTCGTTGGCTTATTGTGAG-3' (reverse for deletion type). DNA concentration was measured using a NanoDrop ND-1000 spectrophotometer (NanoDrop Technologies, Delaware, USA). Reaction mixtures contained 0.5 unit Ex Taq Hot Start Version (Takara), 1.6 µl dNTP mixture (2.5 mM of each dNTP), 2 µl 10X Ex Taq Buffer, 1 µl primer mixture (10 pmol of each primer), and 20 ng genomic DNA in a total volume of 20 µl. The PCR for amplifying the *PKD1L2* region was performed

with an initial step at 98 ℃ for 1 min, followed by 40 cycles of 98℃ for 10 s, 55℃ for 30 s, and 72℃ for 60 s. All PCR products were analyzed on 1% agarose gels.

RNA extraction and cDNA synthesis

Total RNA from the colorectal cancer cell line SNU-C2A (Korean Cell Line Bank, Seoul, Korea), which has a heterozygous deletion of *PKD1L2*, was extracted using Trizol reagent (Invitrogen) according to the manufacturer's instructions. First-strand cDNA was synthesized from 1 μg of total RNA isolated from SNU-C2A and human testis RNA (Clontech, California, USA), which has a normal copy number (CN) of *PKD1L2*, using SuperScript III Reverse transcriptase (Invitrogen) according to the manufacturer's protocol.

PCR amplification of cDNA

Amplification of cDNA using nested PCR was performed to find the *PKD1L2* breakpoint. The PCR primer set targeting exons 26-39 (1^{st} PCR) was 5'-AGAACTGCGGAGCCTCCGGCTG-3' (forward) and 5'-AGGACAGCCCTCTTCACAAAC-3' (reverse) (68), and the primer set targeting exons 28-34 (2^{nd} PCR) was 5'-GACACATCTGGTATTCGATCTTCA-3' (forward) and 5'-CAATGGCAGCGGTGAGAG-3' (reverse). PCR mixtures contained 1.25 unit Ex Taq (Takara), 4 μl dNTP mixture (2.5 mM of each dNTP), 5 μl 10X Ex Taq

Buffer, 1 µl primer mixture (10 pmol of each primer), and 1 µl PCR product in a total volume of 50 µl. The PCR was performed with an initial step at 95℃ for 3 min, followed by 30 cycles of 95℃ for 30 s, 55℃ for 30 s, and 72℃ for 2 min (1^{st} PCR) or 1 min (2^{nd} PCR).

Linkage disequilibrium analysis

A linkage disequilibrium (LD) block was estimated near the candidate CNV regions by the Haploview 4.2 program (69) using SNP genotyping data from 60 CRC patients including 36 case samples of aCGH previously genotyped using an Illumina 610k chip (Illumina, CA, USA) and included in this validation study subjects (70). The LD structure was also screened for in other populations using the international HapMap 3 genotyping data (71), and we evaluated the relationship between of our CNV and nearby tagging SNPs.

Statistical analysis of the *PKD1L2* CNV

Significance of the CNV frequency difference between 2088 controls and 1874 cases for validation experiment estimated using logistic regression model with adjustment for age and sex. *PKD1L2* variant associated with body mass index (BMI). The body mass index (BMI) of each sample was estimated by weight (kg)/height (m2). The significance of the CNV frequency difference between cases and controls was estimated using logistic regression model with adjustment for age and sex. A P value

less than 0.05 was considered to be statistically significant. Statistical analyses were performed using IBM SPSS statistics 21 package software (IBM SPSS, Illinois, USA).

Survival analysis of the *PKD1L2* CNV

Among 1,874 CRC cases, medical records of 743 cases were reviewed. The cumulative hazard of CRC-associated death was determined using the Kaplan–Meier method. Differences between survival curves were tested using the Log rank test. The hazard ratio (HR) and 95% confidence intervals (CI) of the CRC-associated death was computed using the Cox proportional hazard model. To examine potential confounders, multivariate models were adjusted for age, sex, family history of CRC, AJCC stage (I, II, and III), and chemotherapy (none, neo-adjuvant, and adjuvant). According to the recommendations of the Western Pacific regional office of the World Health Organization (72), obesity was defined as body mass index (BMI) $\geq$ 25 kg/m^2 and subgroup analysis was done. Statistical analysis was performed using SPSS version 12.0K (SPSS Inc., Chicago, IL, USA).

Table 10. Characteristics of CRC cases and controls.

Variables		Cases (n=1874)	Controls (n=2088)	p value
Age (years ± SD)		59.8 ± 11.4	48.4 ± 9.0	<0.001*
BMI(kg/m^2± SD)***		23.5±2.9	24.5±2.9	<0.001*
Sex	Males	1170 (62.4%)	1585 (75.9%)	
	Females	704 (37.6%)	503 (24.1%)	<0.001**

*p-value was calculated by the t-test.

**p value was calculated by the chi-square test.

***BMI was evaluated in 1794 patients and 1989 controls

Table 11. Summary of study subjects.

Study	Samples (n)	Male n (%)	Female n (%)	Age mean ± SD
Discovery	Cases (36)	17 (47%)	19 (53%)	39.72 ± 7.47
	Controls (47)	22 (47%)	25 (53%)	36.19 ± 13.95
Replication	Cases (1,838)	1,153 (63%)	685 (37%)	60.16 ± 11.07
	Controls (2,041)	1,563 (77%)	478 (23%)	48.66 ± 8.68
Combined	Cases (1,874)	1,170 (62%)	704 (38%)	59.76 ± 11.36
	Controls (2,088)	1,585 (76%)	503 (24%)	48.38 ± 9.024

SD, standard deviation

Table 12. Nine CNV regions identified in the discovery study using a CGH.

Chr	Start	Stop	Size	CNV	Odds ratio (Number of CNV in CRCs/Controls)	Gene* (part of gene)
Chr2	24,418	205,163	180,746	Gain	Only in CRC (6/0)	*FAM110C* (intron)
Chr5	78,145,544	78,147,588	2,045	Loss	Only in CRC (3/0)	*ARSB* (3' UTR)
Chr9	132,070,319	132,074,244	3,926	Loss	Only in CRC (3/0)	
Chr11	3,795,319	3,796,355	1,037	Gain	Only in CRC (4/0)	*FRAG1* (intron)
Chr1	187,592,455	187,810,947	218,493	Loss	5.75 (4/1)	
Chr7	144,547,893	144,552,899	5,007	Loss	4.50 (6/2)	
Chr13	107,744,862	107,749,881	5,020	Loss	4.18 (3/1)	*TNFSF13B* (intron)
Chr14	25,660,835	25,662,318	1,484	Loss	3.54 (7/3)	
Chr16	79,727,682	79,734,573	6,892	Loss	3.63 (5/2)	*PKD1L2* (CDS)

*Annotated by the RefSeq gene set.

CNV, copy number variation; aCGH, array comparative genomic hybridization

Results

Discovery of CRC susceptibility regions

Nine Candidate CNV regions were selected according to the criteria described above in the methods section (Table 11). Most of the CNVs were CN losses and were located in intergenic or intronic regions. Of the CNVs evaluated, only one was located in the coding region of the *PKD1L2* gene. We focused on this CNV in the subsequent validation study. This CN loss in *PKD1L2* was found in 5 of 36 CRC samples (13.9%) and 2 of 47 normal samples (4.3%).

Identification of deletion region by breakpoint PCR

The estimated size, based on the result of aCGH, of the CNV in *PKD1L2* was 6,892 bp (Table 12). We also used breakpoint PCR to determine the exact region of the genomic deletion, which resulted in a 6,218bp loss in *PKD1L2* (chr16: 79,728,592-79,734,809) (Figure 1). *PKD1L2* is located on human chromosome 16q23.2 and consists of 43 exons. The region encompassing the deletion contains exon 31 (229 bp), exon 32 (205 bp) and part of exon 33 (48 of 97 bp) of *PKD1L2*.

All 83 aCGH samples were re-genotyped for the *PKD1L2* loss region using breakpoint PCR. We found two more deletion samples in CRC samples, and one

sample from among the controls, which had been originally called a deletion in the previous aCGH experiment, was determined to be wild type. These changes made the results more significant compared to those from the previous analysis (OR=10.32, 95% CI 1.20-88.99, *P*=.034) (Table 13).

The altered mRNA structure in the *PKD1L2* deletion

When genomic DNA from eight Korean CRC cell lines was screened, the same deletion in *PKD1L2* was found from one cell line, and was designated SNU-C2A (data not shown). Analysis of the mRNA structure revealed that the CN loss in *PKD1L2* results in an in-frame type deletion in the transcript that skips exons 31-33 (Figure 2). The deleted region is composed of 177 amino acids, and includes two PKD1L2 transmembrane domains (TM4 and TM5) (Figure 3).

Replication study for copy number loss of *PKD1L2*

To further confirm the association between CN loss of *PKD1L2* and CRC, we performed PCR on 1,838 additional cases and 2,041 controls. Of these, 115 patient (6.3%) and 98 control (4.8%) samples were found to have CN loss of the *PKD1L2* gene, thereby demonstrating again a significant association with CRC risk (OR=1.306, 95% CI 0.937-1.821, *P*=.115) (Table 13). The combined analyses of the discovery and replication samples showed a significant *P* value of 0.028 (OR =1.436,

95% CI 1.040-1.983) (Table 13). In addition, we examined whether this association between CN loss and CRC is affected by age and sex, and identified that *PKD1L2* CN loss with age younger than 50 years old is associated with an increased CRC risk (OR=2.138, 95% CI 1.387-3.297, $P=5.8\times10^{-4}$) and identified that high frequencies of *PKD1L2* CN loss were found in male cases (OR=1.519, 95% CI 1.031-2.236, *P*=.034).

Sub-group analysis by BMI

It has been previously reported that *PKD1L2* may be associated with adipogenesis and obesity (73, 74), so we performed a subgroup analysis accoring to BMI (Figure 4). Among 1,874 CRC cases and 2,041 controls, BMI data records were available in 1794 cases and 1989 controls (Figure 4). The CN loss of *PKD1L2* with BMI above 25 exhibited a significant increase in CRC risk (OR=2.287, 95% CI 1.292-4.049, *P*=.005). Especially, we found that *PKD1L2* CNV was significantly associated with significantly increased CRC risk in subjects with the age below 50 and BMI above or equal to 25 (OR=5.238, 95% CI 2.358-11.637, $P= 4.8\times10^{-5}$).

Effect of the *PKD1L2* CNV on the CRC-associated death

The baseline clinical characteristics of the 743 patients enrolled in survival analysis are shown in Table 14. The mean follow-up duration was 60.1 ± 28.1 months and 136

CRC-related deaths were identified. The mean number of surveillance colonoscopies during the study period was 1.2 ± 0.6. The CNV of *PKD1L2* was normal in 689 (92.7%) patients and 54 (7.3%) patients, respectively. There was no significant difference between the patients with normal and abnormal CNV of *PKD1L2*. However, in obese CRC patients, abnormal CNV of *PKD1L2* was associated poor survival rate (*p*=.026, Figure 5). On cox regression model, *PKD1L2* CNV was associated with CRC-associated death significantly only in obese CRC patients (HR=2.96, 95% C, 1.23-7.16, *P*=.016, Table 15).

Table 13. Association of the *PKD1L2* CNV with CRC in cases and controls.

Study	case group	case	control	OR (95% CI)*	*P**
Discovery	all	7/36	1/47	10.32 (1.197-88.990)	**.034**
Replication	all	115/1838	98/2041	1.306 (0.937-1.821)	.115
Combined	all	122/1874	99/2088	1.436 (1.040-1.983)	**.028**
	age < 50	33/364		2.138 (1.387-3.297)	**$5.8x10^{-4}$**
	age ≥ 50	89/1510		0.972 (0.637-1.483)	.894
	Man	73/1170		1.519 (1.031-2.236)	**.034**
	Woman	49/704		1.326 (0.740-2.378)	.344

* OR and *P* value was estimated by logistic regression model with adjustment for age and sex.

Bold characters indicate a statistically significant difference with *P*-values < 0.05.

CNV, copy number variation; CRC, colorectal cancer; OR, odds ratio.

Table 14. Patient characteristics according to the body mass index.

	Total (n=743)	Body mass index		*p*
		< 25 kg/m^2 (n=506)	≥ 25 kg/m^2 (n=237)	
Age (years), mean ± SD	58.8 ± 11.2	60.2 ± 11.7	58.9 ± 10.2	.123
Sex				.935
Male, n	473	323	150	
Female, n	270	183	87	
Family history of CRC				<0.001
Yes, n	74	55	19	
No, n	669	451	218	
AJCC stage				.151
I, n	137	82	55	
II, n	566	396	170	
III, n	40	28	12	
Chemotherapy				.701
None, n	295	197	98	
Neoadjuvant chemotherapy, n	404	277	127	
Adjuvant chemotherapy, n	44	32	12	
CRC associated death				.222
Yes, n	136	99	37	
No, n	607	407	200	
PKD1L2 CNV				.449
Normal	689	472	217	
Abnormal	54	34	20	

Table 15. Association of the *PKD1L2* CNV with CRC-associated death

	All CRC patients (n=743)					Non-obese CRC patients (n=506)					Obese CRC patients (n=237)				
	Death	Univariate analysis		Multivariate analysis*		Death	Univariate analysis		Multivariate analysis*		Death	Univariate analysis		Multivariate analysis*	
CNV of *PKD1L2*	- / +, n	HR (95%CI)	*p*	HR (95%CI)	*P*	- / +, n	HR (95%CI)	*P*	HR (95%CI)	*p*	- / +, n	HR (95%CI)	*p*	HR (95%CI)	*p*
-	122 / 567	-	.286	-	.447	92 / 380	-	.918	-	.947	30 / 187	-	.031	-	.016
+	14 / 40	1.35 (0.78-2.35)		1.25 (0.66-2.22)		7 / 27	0.96 (0.45-2.07)		1.03 (0.47-2.25)		7 / 13	2.49 (1.09-5.68)		2.96 (1.23-7.16)	

CRC, colorectal cancer; CNV, copy number variation.

* adjustment with age, sex, family history of colorectal cancer, AJCC stage, and chemotherapy

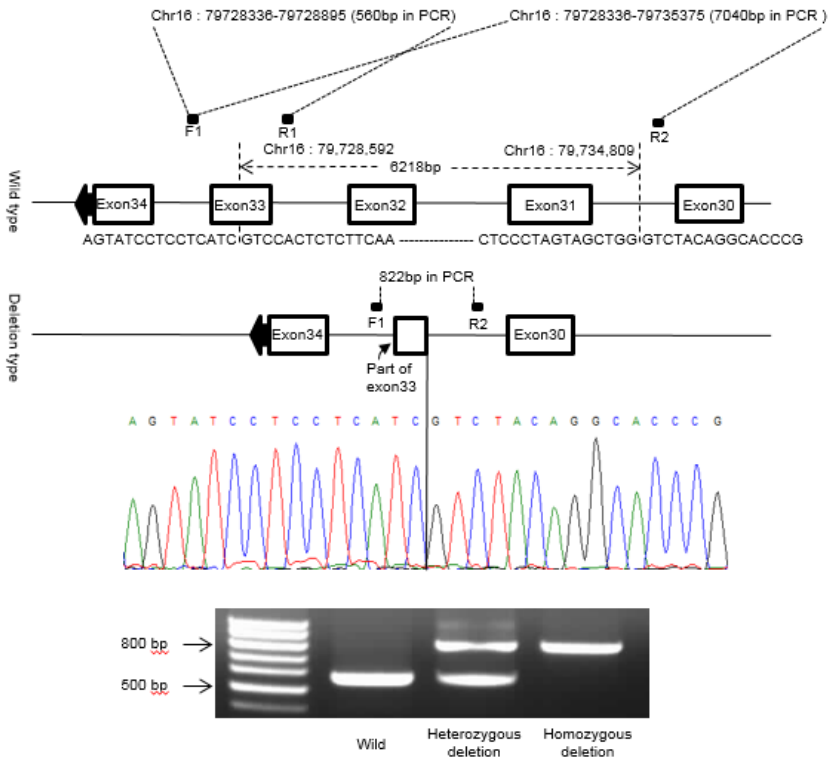


Figure 1. CNV breakpoint sequence analysis of *PKD1L2* loss in DNA. (A) The primer pair of F1 (forward primer) and R1 (reverse primer for normal type) was designed to genotype the normal copy number sequence (560bp) and also to confirm the success of the PCR reaction. The primer pair of F1 and R2 (reverse primer for deletion type) targets copy number loss of *PKD1L2* (822bp) and does not amplify wild-type samples (7040bp), due to the short extension time of the PCR reaction. (B) Electrophoretic analysis of PCR products from wild-type, heterozygous deletion and homozygous deletion samples

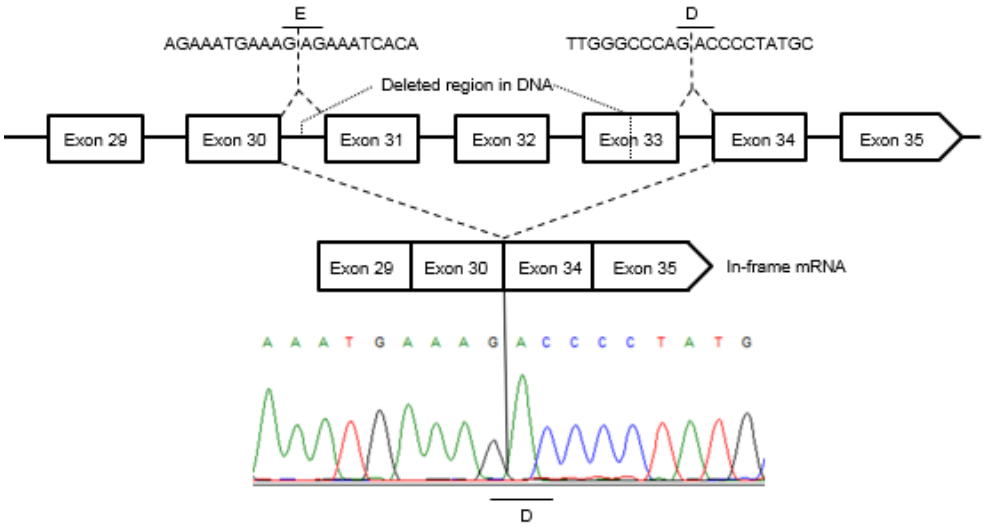


Figure 2. CNV breakpoint sequence analysis of *PKD1L2* loss in RNA. Three exons (531 bases encoding 177 amino acids) were skipped in *PKD1L2* mRNA when the genomic DNA of *PKD1L2* was deleted by the CNV. This exon skipping leads to an in-frame deletion in the *PKD1L2* mRNA.

E: Glutamic acid; D: Aspartic acid. CNV, copy number variation.

Figure 3. Amino acid sequence and domains of PKD1L2. Several TM regions in PKD1L2 are underlined. The deleted region in *PKD1L2* is marked by the grey dotted box. The deleted region is composed of 177 amino acids and includes the TM4 and TM5 domains.

TM, transmembrane.

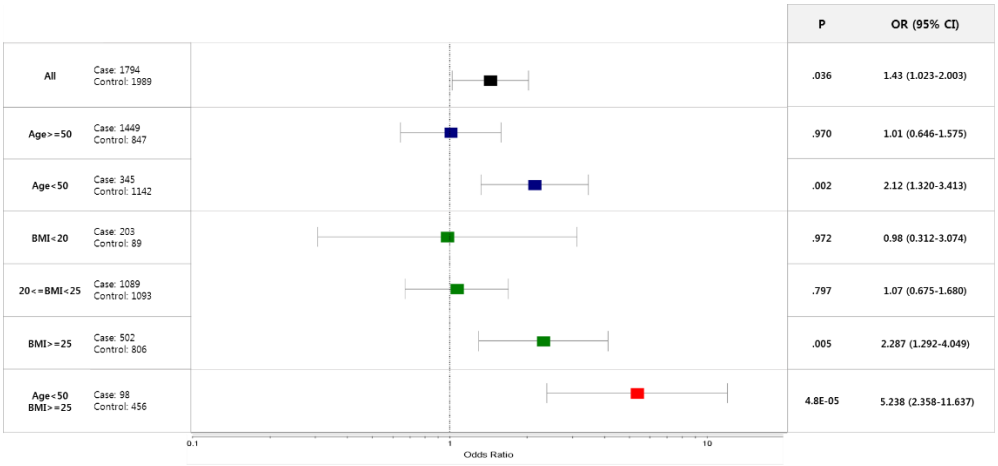


Figure 4. Forest plot for the association between *PKD1L2* CNV and CRC risk stratified by body mass index (BMI) and age. Data for BMI and age were available from 1794 CRC patients and 1989 controls. We estimate the association of *PKD1L2* CNV and CRC risk stratified by BMI and age. OR and *P* value was estimated by logistic regression model with adjustment for age and sex.

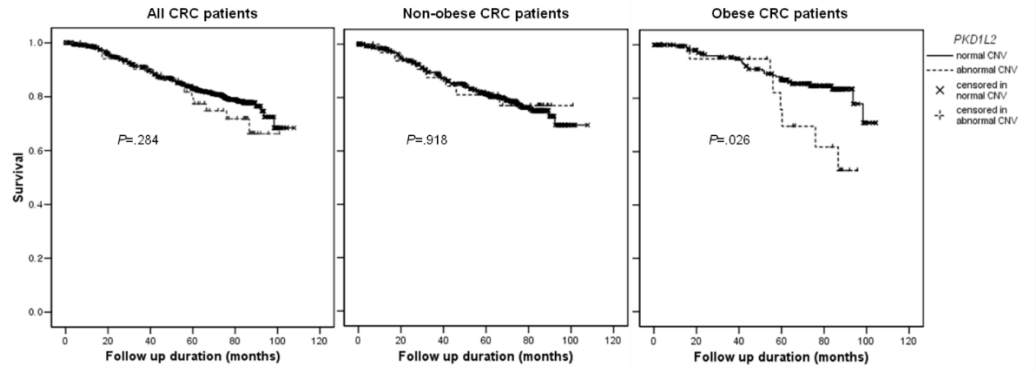


Figure 5. Survival rate of all CRC patients, non-obese CRC, and obese CRC patients according to the CNV of *PKD1L2*. Clinical data were available from 743 CRC patients.

Discussion

In this study, we found an association between a 6.2 kb deletion in the *PKD1L2* gene and CRC in a Korea population using 180k aCGH designed for detection of common Asian-specific CNVs (64). We further validated these findings in a follow-up study of 1,838 cases and 2,041 controls.

The deletion in *PKD1L2* was first found in 2010 in one (NA18582 from Chinese origin) of 30 Asian individuals genotyped by high resolution aCGH (64). Here we determined the exact breakpoints of the deletion and estimated its allele frequency in a Korean population. We also found a similar allele frequency for the *PKD1L2* deletion in a Mongolian population using breakpoint PCR (4/181, allele frequency=0.011). The 1000 Genomes Project reported deletions at almost the same position in four Chinese (NA18621, NA18534, NA18582 and NA18574) and one Japanese (NA19055) individual (75). However, there has been no report in the Database of Genomic Variants (DGV) of the deletion in ethnic groups other than in East Asian populations, thereby suggesting that it might be an East Asian specific CNV.

Our data showed that 6.5% (122 of 1,874 of Korean CRC patients had the deletion in their genomes. The OR was about 1.44 and was slightly higher than those of SNPs reported in other GWAS studies (9, 26, 46-48). Up to date, only a few common CNVs that predispose to CRC have been reported. These include CNVs located in 11q11 (*P*

= .039, OR = 1.122) and the phase II detoxification genes (*P* =.044, OR = 0.82) which were discovered in a European population (62, 63). Since the control group in our study was not a group of disease-free individuals but a general population, the true OR may be higher than 1.44. When we consider only the early onset CRC patients (diagnosed before 50 years old), 9.0% (33 of 364) of the patients had the deletion and the OR was 2.14 (Table 13). Furthermore, when we consider that CRC incidence rates were higher in men than in women (76), our result also showed a more significant association between men and CRC risk (Table 13). These early onset CRC cases and cases seen in men could be suggestive of a hereditary predisposition (76, 77).

PKD1L2 is a member of the polycystin-1-like subfamily (68, 78). Polycystin-1 subfamily proteins have 11 transmembrane domains (TM1-TM11) in addition to REJ, lectin C-type, GPS and LH2/PLAT domains (78, 79). Proteins in this subfamily may function as G-protein coupled receptors that mediate cation channels (68, 80). PKD1L2 may function as an ion channel pore component or may regulate ion channels (78). We showed that the 6.2 kb deletion covering two and half exons (exons 31, 32 and 33) in *PKD1L2* results in a 531 base in-frame deletion in the mRNA and a 177 amino acid deletion (about 17 kDa) spanning two transmembrane domains (TM4 and TM5) (Figure 3).

Since there have been very few reports on the function of PKD1L2, it is difficult to discern the likely effects of this deletion in the pathogenesis of CRC. Expression of *PKD1L2* in normal or cancer cells from the colon has been reported in public

databases such as the Illumina Body Map RNA-seq (www.illumina.com; ArrayExpress ID: E-MTAB-513) and TCGA databases (81). The expression of mutant *PKD1L2* in colon epithelial cells may alter signaling pathways or ion channel activity, thereby increasing the chance of transformation by other environmental stimuli.

Systemic change induced by the expression of *PKD1L2* in other organs may affect the susceptibility to carcinogenesis. It has been reported that up-regulation of *PKD1L2* in a mouse model induced myopathy (82). Interestingly, they reported that the overexpression of *PKD1L2* caused the alteration of fatty acid metabolism, including decreases in triglycerides, free fatty acids and total body fat (82). Some SNPs in the 16q23-24 region of chromosome 16, which includes *PKD1L2*, were also shown be associated with high-density lipoprotein metabolism (83). Linkages with blood lipid levels (84, 85) and adiposity-related phenotypes (74, 84) have also been reported in this region. Since genetic variation in some genes related to fatty acid metabolism, such as the hepatic lipase gene, are associated with CRC development (86), the altered function of *PKD1L2* in lipid metabolism may play a role in the pathogenesis of colorectal cancer. In subgroup analysis according to BMI (body mass index), we found that the *PKD1L2* CN loss is associated with obese individuals exhibited a significant increased risk of CRC. In addition, *PKD1L2* CNV with obese (BMI>=25) and age below 50 is associated with a high risk for colorectal cancer. In survival analysis, CN loss of *PKD1L2* was associated with poor survival only in obese CRC patients, but not in non-obese CRC patients. Obese CRC patients appear to have

worse overall survival than normal-weight patients with CRC. The CNV of *PKD1L2* may play a role in the associations of obesity with survival in CRC patients.

There have been numerous papers linking chr16q23.2 to prostate cancer (87-89). Considering that both prostate cancer and CRC are widely believed to be associated with lipid metabolism (80, 90), it would be reasonable to evaluate the possible association of *PKD1L2* CNV with prostate cancer.

Although we showed a significant association between *PKD1L2* CNV and CRC, GWAS with common SNPs have failed to detect an association between the CNV in this region and CRC, even in Asian populations (91). That may be due to the low allelic frequency of the CNV or the weak LD structure in this region. Although many common CNVs are likely to be tagged by nearby SNPs (92, 93), our data showed no LD between the CNV and the common SNPs in *PKD1L2* (Figure 5). We also estimated LD between the CNV region and neighboring SNPs in other population-based HapMap3 data (Figure 6), and found that it might be difficult to discover significant copy number variations in GWAS studies that used SNP microarrays to genotype samples. In addition, GWASs have failed to detect any association between prostate cancer and this region, despite numerous reports of linkage. Some of these inconsistencies between association and linkage may be explained by structural variations which have little or no LD with common tagging SNPs.

There are several limitations to our study. As we have focused only on Korean

individuals, further studies will be needed in different Asian populations as well as in other ethnicities, including Caucasians and Africans, to confirm our findings. Functional validation will also be required in order to reveal the role of *PKD1L2* in carcinogenesis. There has been one report on the consequences of up-regulation of *PKD1L2* in a mouse model (82), but no reports have evaluated loss of function. Knockout studies in an animal model, such as mice or zebrafish, would give us a clue as to the systemic or local effect of *PKD1L2* deletion.

Despite these limitations, our study has some important implications for understanding CRC. To our knowledge, this is the first report on the association between a common exonal deletion and CRC identified by a genome wide scan. Our results also provide evidence that an association study using a CNV which is not in LD with any of the common SNPs is capable of identifying new disease-related genes. In addition, our study supports the idea that using early onset cases of complex diseases, like CRC, that can be affected by both genetic and environmental factors can increase the likelihood of disease gene discovery. This is an important point to consider in future GWAS studies. Finally, our new findings can be applied to CRC screening of adults younger than 50 years, in whom screening is not currently recommended, despite the increasing incidence of CRC in this population (94).

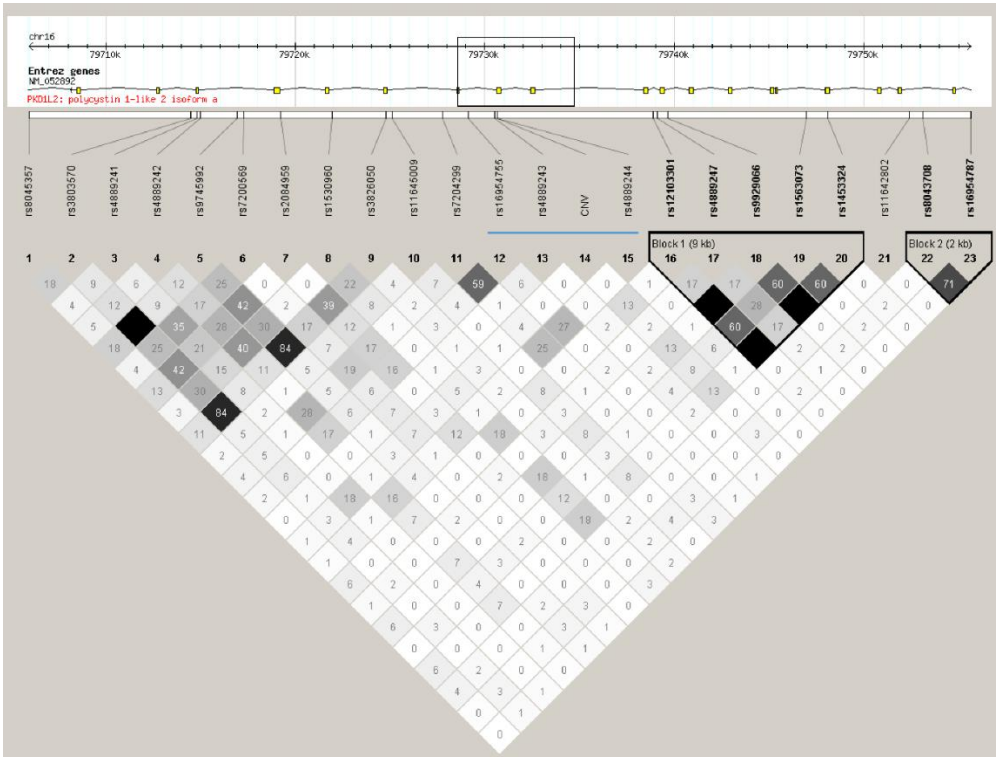
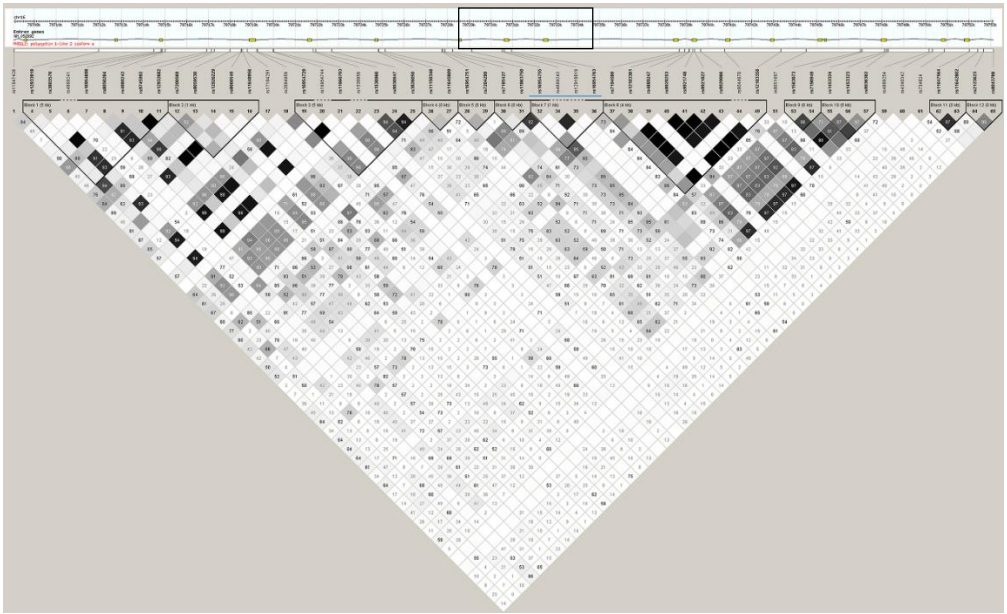


Figure 5. Linkage disequilibrium (LD) in 60 CRC samples. We estimated LD between the *PKD1L2* CNV and neighboring SNPs using an r2 value of 0.8 as the pairwise threshold. SNPs used were located within 50 kb of the *PKD1L2* CNV. The CNV region is depicted inside the box. SNPs in the CNV region are underlined in blue.

CNV, copy number variation; SNP, single nucleotide polymorphism.

A



B

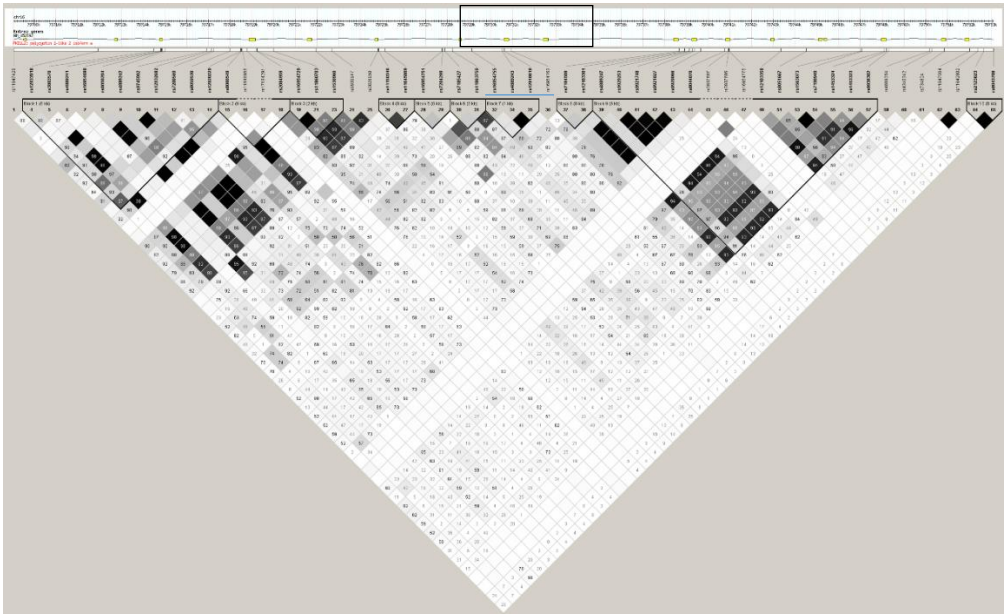


Figure 6. Linkage disequilibrium (LD) between the *PKD1L2* CNV and

neighboring SNPs. Neighboring SNPs were located within 50 kb of the *PKD1L2* CNV. The CNV region is depicted inside the box. SNPs in the CNV region are underlined in blue. (A) LD in CEU+TSI. (B) LD in CHB+JPT

CEU+TSI, Combined panel of Utah residents with Northern and Western European ancestry from the CEPH collection and Tuscans in Italy; JPT+CHB, Combined panel of Japanese in Tokyo, Japan and Han Chinese in Beijing, China; CNV, copy number variation; SNP, single nucleotide polymorphism.

REFERENCES

1. Hong SN, Park C, Kim JI, Kim DH, Kim HC, Chang DK, et al. Colorectal cancer-susceptibility single-nucleotide polymorphisms in Korean population. Journal of gastroenterology and hepatology. 2015;30(5):849-57.

2. Torre LA, Bray F, Siegel RL, Ferlay J, Lortet-Tieulent J, Jemal A. Global cancer statistics, 2012. CA: a cancer journal for clinicians. 2015;65(2):87-108.

3. Lee YS, Choi HB, Lee IK, Kim TG, Oh ST. Association between interleukin-4R and TGF-beta1 gene polymorphisms and the risk of colorectal cancer in a Korean population. Colorectal disease : the official journal of the Association of Coloproctology of Great Britain and Ireland. 2010;12(12):1208-12.

4. Kim DH. Colorectal cancer epidemic in Korea. Annals of coloproctology. 2013;29(1):1-4.

5. Kotake K, Koyama Y, Nasu J, Fukutomi T, Yamaguchi N. Relation of family history of cancer and environmental factors to the risk of colorectal cancer: a case-control study. Japanese journal of clinical oncology. 1995;25(5):195-202.

6. Lichtenstein P, Holm NV, Verkasalo PK, Iliadou A, Kaprio J, Koskenvuo M, et al. Environmental and heritable factors in the causation of cancer--analyses of

cohorts of twins from Sweden, Denmark, and Finland. N Engl J Med. 2000;343(2):78-85.

7. Bost B, de Vienne D, Hospital F, Moreau L, Dillmann C. Genetic and nongenetic bases for the L-shaped distribution of quantitative trait loci effects. Genetics. 2001;157(4):1773-87.

8. Burt RW. Colon cancer screening. Gastroenterology. 2000;119(3):837-53.

9. Houlston RS, Cheadle J, Dobbins SE, Tenesa A, Jones AM, Howarth K, et al. Meta-analysis of three genome-wide association studies identifies susceptibility loci for colorectal cancer at 1q41, 3q26.2, 12q13.13 and 20q13.33. Nat Genet. 2010;42(11):973-7.

10. Aaltonen L, Johns L, Jarvinen H, Mecklin JP, Houlston R. Explaining the familial colorectal cancer risk associated with mismatch repair (MMR)-deficient and MMR-stable tumors. Clinical cancer research : an official journal of the American Association for Cancer Research. 2007;13(1):356-61.

11. Real LM, Ruiz A, Gayan J, Gonzalez-Perez A, Saez ME, Ramirez-Lorca R, et al. A colorectal cancer susceptibility new variant at 4q26 in the Spanish population identified by genome-wide association analysis. PLoS One. 2014;9(6):e101178.

12. Mousses S, Kiefer J, Von Hoff D, Trent J. Using biointelligence to search the cancer genome: an epistemological perspective on knowledge recovery strategies to enable precision medical genomics. Oncogene. 2008;27 Suppl 2:S58-66.

13. Hardy J, Singleton A. Genomewide association studies and human disease. N Engl J Med. 2009;360(17):1759-68.

14. Klein RJ, Zeiss C, Chew EY, Tsai JY, Sackler RS, Haynes C, et al. Complement factor H polymorphism in age-related macular degeneration. Science (New York, NY). 2005;308(5720):385-9.

15. Welter D, MacArthur J, Morales J, Burdett T, Hall P, Junkins H, et al. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. Nucleic acids research. 2014;42(Database issue):D1001-6.

16. Broderick P, Carvajal-Carmona L, Pittman AM, Webb E, Howarth K, Rowan A, et al. A genome-wide association study shows that common alleles of SMAD7 influence colorectal cancer risk. Nat Genet. 2007;39(11):1315-7.

17. Haiman CA, Le Marchand L, Yamamato J, Stram DO, Sheng X, Kolonel LN, et al. A common genetic risk factor for colorectal and prostate cancer. Nat Genet. 2007;39(8):954-6.

18. Jaeger E, Webb E, Howarth K, Carvajal-Carmona L, Rowan A, Broderick P, et al. Common genetic variants at the CRAC1 (HMPS) locus on chromosome 15q13.3 influence colorectal cancer risk. Nat Genet. 2008;40(1):26-8.

19. Tenesa A, Farrington SM, Prendergast JG, Porteous ME, Walker M, Haq N, et al. Genome-wide association scan identifies a colorectal cancer susceptibility locus on 11q23 and replicates risk loci at 8q24 and 18q21. Nat Genet. 2008;40(5):631-7.

20. Tomlinson I, Webb E, Carvajal-Carmona L, Broderick P, Kemp Z, Spain S, et al. A genome-wide association scan of tag SNPs identifies a susceptibility variant for colorectal cancer at 8q24.21. Nat Genet. 2007;39(8):984-8.

21. Tomlinson IP, Webb E, Carvajal-Carmona L, Broderick P, Howarth K, Pittman AM, et al. A genome-wide association study identifies colorectal cancer susceptibility loci on chromosomes 10p14 and 8q23.3. Nat Genet. 2008;40(5):623-30.

22. Zanke BW, Greenwood CM, Rangrej J, Kustra R, Tenesa A, Farrington SM, et al. Genome-wide association scan identifies a colorectal cancer susceptibility locus on chromosome 8q24. Nat Genet. 2007;39(8):989-94.

23. Hinds DA, Stuve LL, Nilsen GB, Halperin E, Eskin E, Ballinger DG, et al. Whole-genome patterns of common DNA variation in three human populations.

Science (New York, NY). 2005;307(5712):1072-9.

24. Adeyemo A, Rotimi C. Genetic variants associated with complex human diseases show wide variation across multiple populations. Public health genomics. 2010;13(2):72-9.

25. Sung JJ, Lau JY, Young GP, Sano Y, Chiu HM, Byeon JS, et al. Asia Pacific consensus recommendations for colorectal cancer screening. Gut. 2008;57(8):1166-76.

26. Houlston RS, Webb E, Broderick P, Pittman AM, Di Bernardo MC, Lubbe S, et al. Meta-analysis of genome-wide association data identifies four new susceptibility loci for colorectal cancer. Nat Genet. 2008;40(12):1426-35.

27. Han DH, Lim SY, Sun BC, Paek DM, Kim HD. Visceral fat area-defined obesity and periodontitis among Koreans. Journal of clinical periodontology. 2010;37(2):172-9.

28. Min BH, Bae JM, Lee EJ, Yu HS, Kim YH, Chang DK, et al. The CpG island methylator phenotype may confer a survival benefit in patients with stage II or III colorectal carcinomas receiving fluoropyrimidine-based adjuvant chemotherapy. BMC cancer. 2011;11:344.

29. Agopian AJ, Eastcott LM, Mitchell LE. Age of onset and effect size in genome-wide association studies. Birth Defects Res A Clin Mol Teratol. 2012;94(11):908-11.

30. Cui R, Okada Y, Jang SG, Ku JL, Park JG, Kamatani Y, et al. Common variant in 6q26-q27 is associated with distal colon cancer in an Asian population. Gut. 2011;60(6):799-805.

31. Matsuo K, Suzuki T, Ito H, Hosono S, Kawase T, Watanabe M, et al. Association between an 8q24 locus and the risk of colorectal cancer in Japanese. BMC cancer. 2009;9:379.

32. Thean LF, Li HH, Teo YY, Koh WP, Yuan JM, Teoh ML, et al. Association of Caucasian-identified variants with colorectal cancer risk in Singapore Chinese. PLoS One. 2012;7(8):e42407.

33. Ho JW, Choi SC, Lee YF, Hui TC, Cherny SS, Garcia-Barcelo MM, et al. Replication study of SNP associations for colorectal cancer in Hong Kong Chinese. Br J Cancer. 2011;104(2):369-75.

34. Xiong F, Wu C, Bi X, Yu D, Huang L, Xu J, et al. Risk of genome-wide association study-identified genetic variants for colorectal cancer in a Chinese population. Cancer Epidemiol Biomarkers Prev. 2010;19(7):1855-61.

35. Li X, Yang XX, Hu NY, Sun JZ, Li FX, Li M. A risk-associated single nucleotide polymorphism of SMAD7 is common to colorectal, gastric, and lung cancers in a Han Chinese population. Mol Biol Rep. 2011;38(8):5093-7.

36. Slattery ML, Herrick J, Curtin K, Samowitz W, Wolff RK, Caan BJ, et al. Increased risk of colon cancer associated with a genetic polymorphism of SMAD7. Cancer Res. 2010;70(4):1479-85.

37. Pittman AM, Webb E, Carvajal-Carmona L, Howarth K, Di Bernardo MC, Broderick P, et al. Refinement of the basis and impact of common 11q23.1 variation to the risk of developing colorectal cancer. Hum Mol Genet. 2008;17(23):3720-7.

38. Halder SK, Beauchamp RD, Datta PK. Smad7 induces tumorigenicity by blocking TGF-beta-induced growth inhibition and apoptosis. Exp Cell Res. 2005;307(1):231-46.

39. Curtin K, Lin WY, George R, Katory M, Shorto J, Cannon-Albright LA, et al. Meta association of colorectal cancer confirms risk alleles at 8q24 and 18q21. Cancer Epidemiol Biomarkers Prev. 2009;18(2):616-21.

40. Xing J, Myers RE, He X, Qu F, Zhou F, Ma X, et al. GWAS-identified colorectal cancer susceptibility locus associates with disease prognosis. Eur J Cancer.

2011.

41. Mates IN, Jinga V, Csiki IE, Mates D, Dinu D, Constantin A, et al. Single nucleotide polymorphisms in colorectal cancer: associations with tumor site and TNM stage. J Gastrointestin Liver Dis. 2012;21(1):45-52.

42. Pomerantz MM, Ahmadiyeh N, Jia L, Herman P, Verzi MP, Doddapaneni H, et al. The 8q24 cancer risk variant rs6983267 shows long-range interaction with MYC in colorectal cancer. Nat Genet. 2009;41(8):882-4.

43. Schmitt AO, Assmus J, Bortfeldt RH, Brockmann GA. CandiSNPer: a web tool for the identification of candidate SNPs for causal variants. Bioinformatics. 2010;26(7):969-70.

44. Mrkonjic M, Roslin NM, Greenwood CM, Raptis S, Pollett A, Laird PW, et al. Specific variants in the MLH1 gene region may drive DNA methylation, loss of protein expression, and MSI-H colorectal cancer. PLoS One. 2010;5(10):e13314.

45. Shia J, Ellis NA, Paty PB, Nash GM, Qin J, Offit K, et al. Value of histopathology in predicting microsatellite instability in hereditary nonpolyposis colorectal cancer and sporadic colorectal cancer. The American journal of surgical pathology. 2003;27(11):1407-17.

46. Peters U, Jiao S, Schumacher FR, Hutter CM, Aragaki AK, Baron JA, et al. Identification of Genetic Susceptibility Loci for Colorectal Tumors in a Genome-Wide Meta-analysis. Gastroenterology. 2013;144(4):799-807.e24.

47. Speicher MR, Geigl JB, Tomlinson IP. Effect of genome-wide association studies, direct-to-consumer genetic testing, and high-speed sequencing technologies on predictive genetic counselling for cancer risk. The lancet oncology. 2010;11(9):890-8.

48. Jia WH, Zhang B, Matsuo K, Shin A, Xiang YB, Jee SH, et al. Genome-wide association analyses in East Asians identify new susceptibility loci for colorectal cancer. Nat Genet. 2013;45(2):191-6.

49. McCarthy MI, Hirschhorn JN. Genome-wide association studies: potential next steps on a genetic journey. Hum Mol Genet. 2008;17(R2):R156-65.

50. Cook EH, Jr., Scherer SW. Copy-number variations associated with neuropsychiatric conditions. Nature. 2008;455(7215):919-23.

51. Girirajan S, Campbell CD, Eichler EE. Human copy number variation and complex genetic disease. Annual review of genetics. 2011;45:203-26.

52. Sebat J, Lakshmi B, Troge J, Alexander J, Young J, Lundin P, et al. Large-

scale copy number polymorphism in the human genome. Science (New York, NY). 2004;305(5683):525-8.

53. Iafrate AJ, Feuk L, Rivera MN, Listewnik ML, Donahoe PK, Qi Y, et al. Detection of large-scale variation in the human genome. Nat Genet. 2004;36(9):949-51.

54. Diskin SJ, Hou C, Glessner JT, Attiyeh EF, Laudenslager M, Bosse K, et al. Copy number variation at 1q21.1 associated with neuroblastoma. Nature. 2009;459(7249):987-91.

55. Aitman TJ, Dong R, Vyse TJ, Norsworthy PJ, Johnson MD, Smith J, et al. Copy number polymorphism in Fcgr3 predisposes to glomerulonephritis in rats and humans. Nature. 2006;439(7078):851-5.

56. Pinto D, Pagnamenta AT, Klei L, Anney R, Merico D, Regan R, et al. Functional impact of global rare copy number variation in autism spectrum disorders. Nature. 2010;466(7304):368-72.

57. St Clair D. Copy number variation and schizophrenia. Schizophrenia bulletin. 2009;35(1):9-12.

58. Levin B, Lieberman DA, McFarland B, Andrews KS, Brooks D, Bond J, et

al. Screening and surveillance for the early detection of colorectal cancer and adenomatous polyps, 2008: a joint guideline from the American Cancer Society, the US Multi-Society Task Force on Colorectal Cancer, and the American College of Radiology. Gastroenterology. 2008;134(5):1570-95.

59. Krepischi AC, Pearson PL, Rosenberg C. Germline copy number variations and cancer predisposition. Future oncology (London, England). 2012;8(4):441-50.

60. Venkatachalam R, Verwiel ET, Kamping EJ, Hoenselaar E, Gorgens H, Schackert HK, et al. Identification of candidate predisposing copy number variants in familial and early-onset colorectal cancer patients. International journal of cancer Journal international du cancer. 2011;129(7):1635-42.

61. de Voer RM, Geurts van Kessel A, Weren RD, Ligtenberg MJ, Smeets D, Fu L, et al. Germline mutations in the spindle assembly checkpoint genes BUB1 and BUB3 are risk factors for colorectal cancer. Gastroenterology. 2013;145(3):544-7.

62. Fernandez-Rozadilla C, Cazier JB, Tomlinson I, Brea-Fernandez A, Lamas MJ, Baiget M, et al. A genome-wide association study on copy-number variation identifies a 11q11 loss as a candidate susceptibility variant for colorectal cancer. Human genetics. 2013.

63. Angstadt AY, Berg A, Zhu J, Miller P, Hartman TJ, Lesko SM, et al. The

effect of copy number variation in the phase II detoxification genes UGT2B17 and UGT2B28 on colorectal cancer risk. Cancer. 2013;119(13):2477-85.

64. Park H, Kim JI, Ju YS, Gokcumen O, Mills RE, Kim S, et al. Discovery of common Asian copy number variants using integrated high-resolution array CGH and massively parallel DNA sequencing. Nat Genet. 2010;42(5):400-5.

65. Ju YS, Hong D, Kim S, Park SS, Kim S, Lee S, et al. Reference-unbiased copy number variant analysis using CGH microarrays. Nucleic acids research. 2010;38(20):e190.

66. Scherer SW, Lee C, Birney E, Altshuler DM, Eichler EE, Carter NP, et al. Challenges and standards in integrating surveys of structural variation. Nat Genet. 2007;39(7 Suppl):S7-15.

67. Rozen S, Skaletsky H. Primer3 on the WWW for general users and for biologist programmers. Methods in molecular biology (Clifton, NJ). 2000;132:365-86.

68. Yuasa T, Takakura A, Denker BM, Venugopal B, Zhou J. Polycystin-1L2 is a novel G-protein-binding protein. Genomics. 2004;84(1):126-38.

69. Barrett JC, Fry B, Maller J, Daly MJ. Haploview: analysis and visualization

of LD and haplotype maps. Bioinformatics. 2005;21(2):263-5.

70. Hong SN, Park CH, Kim JI, Kim DH, Kim HC, Chang DK, et al. Colorectal Cancer-Susceptibility Single Nucleotide Polymorphisms in Korean Population. Journal of gastroenterology and hepatology. 2013.

71. Altshuler DM, Gibbs RA, Peltonen L, Altshuler DM, Gibbs RA, Peltonen L, et al. Integrating common and rare genetic variation in diverse human populations. Nature. 2010;467(7311):52-8.

72. Organization WH. The Asia-Pacific perspective. Redefining obesity and its treatment.

73. Aberg K, Dai F, Sun G, Keighley ED, Indugula SR, Roberts ST, et al. Susceptibility loci for adiposity phenotypes on 8p, 9p, and 16q in American Samoa and Samoa. Obesity (Silver Spring, Md). 2009;17(3):518-24.

74. Dong C, Beecham A, Slifer S, Wang L, McClendon MS, Blanton SH, et al. Genome-wide linkage and peak-wide association study of obesity-related quantitative traits in Caribbean Hispanics. Human genetics. 2011;129(2):209-19.

75. Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE, et al. An integrated map of genetic variation from 1,092 human genomes. Nature.

2012;491(7422):56-65.

76. Haggar FA, Boushey RP. Colorectal cancer epidemiology: incidence, mortality, survival, and risk factors. Clinics in colon and rectal surgery. 2009;22(4):191-7.

77. Giraldez MD, Lopez-Doriga A, Bujanda L, Abuli A, Bessa X, Fernandez-Rozadilla C, et al. Susceptibility genetic variants associated with early-onset colorectal cancer. Carcinogenesis. 2012;33(3):613-9.

78. Li A, Tian X, Sung S-W, Somlo S. Identification of two novel polycystic kidney disease-1-like genes in human and mouse genomes. Genomics. 2003;81(6):596-608.

79. Nims N, Vassmer D, Maser RL. Transmembrane domain analysis of polycystin-1, the product of the polycystic kidney disease-1 (PKD1) gene: evidence for 11 membrane-spanning domains. Biochemistry. 2003;42(44):13035-48.

80. Suburu J, Chen YQ. Lipids and prostate cancer. Prostaglandins & other lipid mediators. 2012;98(1-2):1-10.

81. Network CGA. Comprehensive molecular characterization of human colon and rectal cancer. Nature. 2012;487(7407):330-7.

82. Mackenzie FE, Romero R, Williams D, Gillingwater T, Hilton H, Dick J, et al. Upregulation of PKD1L2 provokes a complex neuromuscular disease in the mouse. Hum Mol Genet. 2009;18(19):3553-66.

83. Dastani Z, Pajukanta P, Marcil M, Rudzicz N, Ruel I, Bailey SD, et al. Fine mapping and association studies of a high-density lipoprotein cholesterol linkage region on chromosome 16 in French-Canadian subjects. European journal of human genetics : EJHG. 2010;18(3):342-7.

84. Dong C, Beecham A, Wang L, Slifer S, Wright CB, Blanton SH, et al. Genetic loci for blood lipid levels identified by linkage and association analyses in Caribbean Hispanics. Journal of lipid research. 2011;52(7):1411-9.

85. Gagnon F, Jarvik GP, Badzioch MD, Motulsky AG, Brunzell JD, Wijsman EM. Genome scan for quantitative trait loci influencing HDL levels: evidence for multilocus inheritance in familial combined hyperlipidemia. Human genetics. 2005;117(5):494-505.

86. Crous-Bou M, Rennert G, Salazar R, Rodriguez-Moranta F, Rennert HS, Lejbkowicz F, et al. Genetic polymorphisms in fatty acid metabolism genes and colorectal cancer. Mutagenesis. 2012;27(2):169-76.

87. Suarez BK, Lin J, Burmester JK, Broman KW, Weber JL, Banerjee TK, et

al. A genome screen of multiplex sibships with prostate cancer. American journal of human genetics. 2000;66(3):933-44.

88. Witte JS, Suarez BK, Thiel B, Lin J, Yu A, Banerjee TK, et al. Genome-wide scan of brothers: replication and fine mapping of prostate cancer susceptibility and aggressiveness loci. The Prostate. 2003;57(4):298-308.

89. Lange EM, Beebe-Dimmer JL, Ray AM, Zuhlke KA, Ellis J, Wang Y, et al. Genome-wide linkage scan for prostate cancer susceptibility from the University of Michigan Prostate Cancer Genetics Project: suggestive evidence for linkage at 16q23. The Prostate. 2009;69(4):385-91.

90. Ong ES, Zou L, Li S, Cheah PY, Eu KW, Ong CN. Metabolic profiling in colorectal cancer reveals signature metabolic shifts during tumorigenesis. Molecular & cellular proteomics : MCP. 2010.

91. Hindorff LA, Junkins HA, Hall PN, Mehta JP, Manolio TA. A catalog of published genome-wide association studies. (Available at: www.genome.gov/gwastudies. Accessed December 20, 2010).

92. Hinds DA, Kloek AP, Jen M, Chen X, Frazer KA. Common deletions and SNPs are in linkage disequilibrium in the human genome. Nat Genet. 2006;38(1):82-5.

93. Craddock N, Hurles ME, Cardin N, Pearson RD, Plagnol V, Robson S, et al. Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls. Nature. 2010;464(7289):713-20.

94. Siegel RL, Jemal A, Ward EM. Increase in incidence of colorectal cancer among young men and women in the United States. Cancer Epidemiol Biomarkers Prev. 2009;18(6):1695-8.

국문초록

한국인에서 대장암과 연관된 유전자 단일염기다형성과 복제수 변이의 확인

서울대학교 대학원 의과학과 의과학 전공

박 창 호

서론: 대장암은 전세계적으로 가장 흔한 암 중 하나이다. 대장암 발생에 기여하는 유전적 변이를 찾기 위해, 전장유전체연관분석연구 (genome-wide association study, GWAS)가 진행되어 왔으며, 최근 서양인에서 전장유전체 연관분석연구를 통해 대장암과 연관된 여러 개의 단일 염기 다형성 (single nucleotide polymorphism, SNP)을 확인하였다. 하지만, 유전적 변이는 인종에 따른 다양성을 가지기 때문에 서양인의 전장유전체여관분석연구 결과로 확인된 단일 염기 다형성이 한국인의 대장암 발생과 관련이 있는지는 불명확하다. 지난 몇 년 동안 대장암의 발생률은 아시아인, 특히 한국인에서 급격히 증가하고 있다. 하지만 한국인의 대장암 감수성 유전자 변이는 명확히 알려져 있지 않다. 따라서, 이번 연구는 한국인 대장암 위

험과 관련된 유전적 변이를 확인하고자 했다. 본 연구의 첫 번째 부분에서는, 대장암과 연관이 있는 단일 염기 다형성 (single nucleotide polymorphism)에 관한 연구이다. 우선, 서양인의 GWAS 에서 대장암과 연관되어 있다고 알려진 SNP 중에 5 개의 SNP 을 선별하여 한국인 대장암과 연관성이 있는지 확인하고자 했으며, 또한 한국인 대장암과 연관된 새로운 SNP 을 전장유전체분석연구 (genome-wide screening)를 이용하여 확인하고자 했다. 이 논문의 두 번째 부분에서는, 최근까지 연구된 GWAS 에서 20 개 이상의 대장암과 관련된 감수성 관련 유전영역이 알려져 있다. 하지만 대부분의 전장연관성연구는 단일 염기 다형성 (SNP)을 이용한 연구이기에, 이들 유전변이는 대장암에 대한 유전성 (heritability)을 극히 일부만 설명할 수 있다. 최근의 여러 논문에서 보고된 바에 따르면 유전자 복제수 변이 (copy number variation)가 암을 비롯한 여러 질병과 연관이 있다는 연구결과가 보고되고 있다. 하지만, 대장암 발생과 진행에 복제수변이의 기여는 명확하게 알려져 있지 않다. 본 연구의 두 번째 부분에서는 비교유전체 보합법 (array comparative genomic hybridization) 방법을 이용하여 대장암과 연관이 있는 유전자 복제수 변이를 확인하고자 하였다.

방법: 서양인에서 보고된 대장암 감수성 유전자들의 인종에 따른 다양성을 확인하기 위하여, 서양인에서 이미 보고된 대장암 감수성 단일 염기 다형성 중 5 개의 유전변이 (rs3802842, rs4779584, rs4939827, rs6983267 과

rs10795668)을 선정하여 한국인 대장암 환자 198 명과 한국인 정상인 329 명을 TaqMan SNP Genotyping Assay 를 이용하여 검증연구 (replication)을 수행하였다. 또한, 대장암과 연관된 새로운 SNP 을 확인하기 위하여 한국인 대장암 환자 45 명과 60 명을 Illumina HumanHap 370K 와 610K BeadChips 을 이용하여 전장 유전체 분석을 하였다. 각각의 유전형 빈도를 한국인 대조군 (the Korean HapMap data)의 유전형 빈도와 비교하였고, 대장암과 연관성이 있는 8 개의 유전변이 후보 (candidate SNPs)를 선정하였다 (rs12266240, rs10491619, rs10941887, rs1859915, rs727235, rs17051076, rs1599695 와 rs902960). 선정된 8 개의 후보 유전변이를 대장암 환자 189 명과 정상인 190 명을 TaqMan SNP Genotyping Assay 를 이용하여 검증연구를 수행하였다. 본 논문의 두 번째 연구부분에서는, 대장암과 연관된 유전자 복제수 변이를 확인하기 위하여 대장암 환자 36 명과 정상인 47 명을 Agilent 사의 180K microaray 를 사용하여 비교유전체 보합법 (array comparative genomic hybridization)을 수행한 후, 대장암과 연관된 후보 유전자 복제수 변이 (candidate CNVs)을 선정하였다. 후보 유전자 복제수 변이 중 한 변이가 *PKD1L2* 유전자의 단백질 암호 부위 (protein-coding region)에서 결실이 존재 하는 것이 확인되었고, 이 복제수 변이의 정확한 결실 중지점 (deletion breakpoint)을 확인하기 위해 중합효소 연쇄반응 (polymerase chain reaction)을 이용하였다. 중합효소 연쇄반응을 이용하여 한

국인 대장암환자 1,874 명과 한국인 대조군 2,088 명을 이용하여 *PKD1L2* 유전자의 복제수 변이가 대장암과 연관이 있는지 확인하였다.

결과: 서양인에서 대장암 발생과 연관이 있다고 알려진 5 개의 단일 염기 다형성 중에, SMAD4 유전자의 rs4939827 은 한국인에서 대장암 위험성을 감소시키는 것으로 확인되었다 [age/gender-adjusted OR (95%CI)]: [additive model, 0.67 (0.47-0.95); dominant model, 0.59 (0.39-0.91)]. Rs10795668 [additive model, 0.61 (0.39-0.95), *P*=.029] 과 rs4779584 [dominant model, 2.50 (1.16-5.39), *P*=.020]은 각각 남성과 여성의 대장암에 연관이 있는 것으로 확인되었다. 임상정보를 이용하여 하위 집단 분석(subgroup analysis)을 하여, rs4939827 이 3/4 기 (stage III/IV status), 림프절 전이 (lymph node metastasis) 또는 원격전이 (distant metastasis)인 진행형 대장암 (advanced CRC)과 연관성이 있음을 확인하였다 (stage III/IV status [additive model, 0.63 (0.43-0.93), *P*=.020], lymph node metastasis [additive model, 0.50 (0.61-0.82), *P*=.006], and distant metastasis [additive model, 0.22 (0.08-0.66), *P*=.007]). 전장 유전체 분석을 통하여 확인된 8 개의 후보 유전변이 중에 rs17051076 은 검증연구에서 고빈도 현미부수체 불안정형 (high microsatellite instability; MSI-H) 대장암과 연관이 있음을 확인하였다 [age/gender-adjusted OR (95%CI), *P*]: additive model, 4.25 (1.51-11.98, *P*=.006); dominant model, 3.52 (1.13-10.94), *P*=.030]. 본 연구의 두 번째 부분은, 비교유전체 보합법 결과 (array comparative genomic hybridization)를 분석하여, 단백질 암호 영역에 복제수 변이를 가지고 있

는 *PKD1L2* 유전자가 대장암과 연관이 있음을 확인하였다. 중합효소 연쇄 반응 (polymerase chain reaction)을 이용하여 *PKD1L2* 유전자의 복제수 변이의 정확한 결손 부위를 확인하였다. 한국인 대장암 환자 1,874 명과 정상인 대조군 2,088 명을 사용하여 검증연구 한 결과 *PKD1L2* 유전자 복제수 변이가 대장암과 유의하게 연관성이 있는 것을 확인하였으며 (OR=1.44, 95% CI=1.04-1.98, P=.028), 또한, 50 세 이하의 집단에서 *PKD1L2* 복제수 변이가 대장암과 연관이 있는 것으로 확인되었다 (OR=2.14, 95% CI 1.39-3.30, P=5.8x10^{-4}). 신체질량지수 (body mass index)에 따른 하위집단 (subgroup) 분석을 통하여, 비만 (BMI >=25)을 가진 한국인에서 *PKD1L2* 복제수 변이와 대장암의 연관이 있음을 확인하였으며(OR=2.29, 95% CI 1.29-4.05, P=.005), 특히 50세 이하의 비만을 가진 그룹에서 *PKD1L2*복제수 변이가 대장암의 발생 위험이 상당히 증가하는 것을 확인하였다 (OR=5.24, 95% CI 2.36-11.64, P= 4.8x10^{-5}). 더욱이, 비만을 가진 대장암 환자그룹에서 *PKD1L2* 의 유전자 복제수 변이가 짧은 생존률과 연관이 있음을 확인하였다 (P=0.026).

결론: 서양인에서 대장암과 연관이 있다고 알려진 rs4939827, rs4779584 와 rs10795668 은 한국인 대장암발생 위험에 영향을 줄 수 있다. 새로 확인한 rs17051076 은 한국인에서 고빈도 현미부수체 불안정형 대장암과 연관이 있다는 것을 확인하였다. *PKD1L2* 유전자의 복제수 변이가 대장암과 연관

이 있음을 확인하였으며, 비만 (BMI >=25)인 50 세 이하의 그룹에서는 *PKD1L2* 의 유전자 복제수 변이가 대장암의 발생위험을 상당히 증가시키는 것으로 확인되었다. 또한, 비만인 대장암 환자에서는 *PKD1L2* 유전자 복제수 변이가 짧은 생존률과 연관이 있다는 것을 확인하였다.

* 본 내용은 Journal of Gastroenterology and Hepatology 에 출판 완료된 내용임 (1)

주요어 : 단일 염기 다형성; 유전자 복제수 변이; 대장암; 전장 연관성 연구; 고빈도 현미부수체 불안정형; PKD1L2, 체질량지수
학 번 : 2010-31168