nature genetics

Genome-wide association analyses of esophageal squamous cell carcinoma in Chinese identify multiple susceptibility loci and gene-environment interactions

Chen Wu[1,2,13], Peter Kraft[2,13], Kan Zhai[1,13], Jiang Chang[1,13], Zhaoming Wang[3,4], Yun Li[5], Zhibin Hu[6], Zhonghu He[7], Weihua Jia[8], Christian C Abnet[3], Liming Liang[2], Nan Hu[3], Xiaoping Miao[9], Yifeng Zhou[10], Zhihua Liu[1], Qimin Zhan[1], Yu Liu[1], Yan Qiao[1], Yuling Zhou[1], Guangfu Jin[6], Chuanhai Guo[7], Changdong Lu[11], Haijun Yang[11], Jianhua Fu[8], Dianke Yu[1], Neal D Freedman[3], Ti Ding[12], Wen Tan[1], Alisa M Goldstein[3], Tangchun Wu[9], Hongbing Shen[6], Yang Ke[7], Yixin Zeng[8], Stephen J Chanock[3,4], Philip R Taylor[3] & Dongxin Lin[1]

We conducted a genome-wide association study (GWAS) and a genome-wide gene-environment interaction analysis of esophageal squamous-cell carcinoma (ESCC) in 2,031 affected individuals (cases) and 2,044 controls with independent validation in 8,092 cases and 8,620 controls. We identified nine new ESCC susceptibility loci, of which seven, at chromosomes 4q23, 16q12.1, 17q21, 22q12, 3q27, 17p13 and 18p11, had a significant marginal effect ($P = 1.78 \times 10^{-39}$ to $P = 2.49 \times 10^{-11}$) and two of which, at 2q22 and 13q33, had a significant association only in the gene–alcohol drinking interaction (gene-environment interaction P ($P_{G \times E}$) $= 4.39 \times 10^{-11}$ and $P_{G \times E} = 4.80 \times 10^{-8}$, respectively). Variants at the 4q23 locus, which includes the *ADH* cluster, each had a significant interaction with alcohol drinking in their association with ESCC risk ($P_{G \times E} = 2.54 \times 10^{-7}$ to $P_{G \times E} = 3.23 \times 10^{-2}$). We confirmed the known association of the *ALDH2* locus on 12q24 to ESCC, and a joint analysis showed that drinkers with both of the *ADH1B* and *ALDH2* risk alleles had a fourfold increased risk for ESCC compared to drinkers without these risk alleles. Our results underscore the direct genetic contribution to ESCC risk, as well as the genetic contribution to ESCC through interaction with alcohol consumption.

ESCC ranks as the tenth most prevalent cancer in the world, with marked regional variation and a particularly high incidence in certain regions of China. Previous molecular epidemiological studies using a candidate gene approach have implicated a set of genetic variations that confer susceptibility to ESCC, primarily variations that are related to alcohol metabolism[1–6]. The GWAS has emerged as a powerful and successful tool to identify common disease alleles by using high-throughput genotyping technology to interrogate a large number of tagging SNPs that serve as surrogates for untested common SNPs across the genome. In studies published thus far, GWAS of cancers of the upper aerodigestive tract, including ESCC in individuals of European[7,8] and Japanese ancestry[9], have shown that variants in *ADH* genes and/or *ALDH2* are associated with risk of ESCC; in addition, these studies have shown an interaction for these loci with alcohol. Two GWAS showed that variants in *PLCE1* and, perhaps, *C20orf54* are associated with risk of ESCC in Chinese populations[10,11].

We recently reported a multistage GWAS of ESCC that was based on genotyping 666,141 SNPs in 2,031 cases and 2,044 controls with a second replication stage in 6,276 cases and 6,165 controls and identified three new loci that are associated with susceptibility to ESCC[12]. In this previous study, we attempted to replicate 29 SNPs with $P \leq 10^{-7}$. Because of our use of this stringent P value threshold, it is possible that some true ESCC-associated loci with moderate effect sizes were overlooked[13]. However, such loci may be detected by dense genotyping or analyzing larger sample sizes[14]. Furthermore, in our published GWAS report, we observed that three variants at 12q24 conferred ESCC risk through a gene-lifestyle interaction, with a pronounced elevation of risk among alcohol users[12]. Alcohol intake is an important risk factor that contributes to the development of ESCC in Asian and other populations[15]. These findings underscore the fact that ESCC is a complex disease and that its etiology is related to environmental exposures, multiple genetic loci and gene-environment interactions.

[1]State Key Laboratory of Molecular Oncology, Cancer Institute and Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing, China. [2]Department of Epidemiology, Harvard School of Public Health, Boston, Massachusetts, USA. [3]Division of Cancer Epidemiology and Genetics, National Cancer Institute (NCI), US National Institutes of Health, Bethesda, Maryland, USA. [4]Core Genotyping Facility, NCI-Frederick, SAIC-Frederick, Frederick, Maryland, USA. [5]Department of Genetics, University of North Carolina, Chapel Hill, North Carolina, USA. [6]Department of Epidemiology and Biostatistics, Cancer Center, Nanjing Medical University, Nanjing, Jiangsu, China. [7]Key Laboratory of Carcinogenesis and Translational Research (Ministry of Education), Peking University School of Oncology, Beijing Cancer Hospital and Institute, Beijing, China. [8]State Key Laboratory of Oncology in Southern China, Sun Yat-Sen University Cancer Center, Guangzhou, Guangdong, China. [9]Key Laboratory for Environment and Health (Ministry of Education), School of Public Health, Huazhong University of Sciences and Technology, Wuhan, Hubei, China. [10]Laboratory of Cancer Molecular Genetics, Medical College of Soochow University, Suzhou, Jiangsu, China. [11]Anyang Cancer Hospital, Anyang, Henan, China. [12]Shanxi Cancer Hospital, Taiyuan, Shanxi, China. [13]These authors contributed equally to this work. Correspondence should be addressed to D.L. (lindx72@cicams.ac.cn).

Received 5 March; accepted 16 August; published online 9 September 2012; doi:10.1038/ng.2411

Table 1 Characteristics of cases with ESCC and controls who participated in this study

	GWAS[a]		Replication 1[a]		Replication 2[b]		Combined sample		High-risk cohort[c]	
	Cases (*N* = 2,031)	Controls (*N* = 2,044)	Cases (*N* = 3,571)	Controls (*N* = 3,602)	Cases (*N* = 4,521)	Controls (*N* = 5,018)	Cases (*N* = 10,123)	Controls (*N* = 10,664)	Cases (*N* = 1,410)	Controls (*N* = 1,656)
Age, mean (s.d.)	59.8 (9.8)	61.3 (8.5)	60.5 (8.9)	55.7 (12.7)	60.1 (9.0)	51.5 (13.5)	60.2 (9.1)	54.8 (12.9)	58.1 (8.1)	57.8 (9.2)
Sex										
Male, *N* (%)	1,627 (80.1)	1,706 (83.5)	2,653 (74.3)	2,374 (65.9)	3,380 (74.8)	3,955 (78.8)	7,660 (75.7)	8,035 (75.3)	919 (65.2)	1,222 (73.8)
Female, *N* (%)	404 (19.9)	338 (16.5)	918 (25.7)	1,228 (34.1)	1,141 (25.2)	1,063 (21.2)	2,463 (24.3)	2,629 (24.7)	491 (34.8)	434 (26.2)
Smoking status										
Nonsmoker, *N* (%)	706 (34.8)	895 (43.8)	1,604 (44.9)	2,082 (57.8)	2,115 (46.8)	2,542 (50.7)	4,425 (43.7)	5,519 (51.8)	551 (39.1)	577 (34.8)
Smoker, *N* (%)	1,325 (65.2)	1,149 (56.2)	1,967 (55.1)	1,520 (42.2)	2,406 (53.2)	2,476 (49.3)	5,698 (56.3)	5,145 (48.2)	859 (60.9)	1,079 (65.2)
Drinking status										
Nondrinker, *N* (%)	886 (43.6)	1,139 (55.7)	1,982 (55.5)	2,307 (64.0)	2,115 (46.8)	2,886 (57.5)	4,983 (49.2)	6,332 (59.4)	1,112 (78.9)	1,373 (82.9)
Drinker, *N* (%)	1,145 (56.4)	905 (44.3)	1,589 (44.5)	1,295 (36.0)	2,406 (53.2)	2,132 (42.5)	5,140 (50.8)	4,332 (40.6)	298 (21.1)	283 (17.1)

[a]Cases and controls were recruited from Beijing region. [b]Cases and controls were recruited from Jiangsu, Henan and Guangdong provinces. [c]This case-control set, derived from Shanxi province, where the ESCC incidence and mortality rates are among the highest in China[11], had considerably lower percentage of drinkers in both the case and control categories compared with the percentages in other groups.

Because some ESCC susceptibility loci act in an environment-responsive manner, true associations might not be detected by GWAS without accounting for environmental risk factors[16]. Thus, to discover these susceptibility loci, incorporation of environmental risk factors in the context of GWAS may yield additional regions that are worthy of follow-up studies.

Here we report a new, multistage GWAS of ESCC in a total of 10,123 cases with ESCC and 10,664 controls (**Table 1**). We also report, to our knowledge, the first genome-wide gene-environment interaction analysis of ESCC that incorporates alcohol drinking. We replicated results from these GWAS in an additional case-control panel from a high-risk population.

RESULTS

New loci associated with susceptibility to ESCC

To identify new susceptibility loci for ESCC, we analyzed 169 promising SNPs (with $10^{-7} < P < 10^{-4}$ in our previous GWAS; **Supplementary Table 1**) in replication 1 comprising 3,571 cases and 3,602 controls. We further verified the 18 SNPs with $P < 0.01$ in replication 2 comprising 4,521 cases and 5,018 controls. We found that 15 SNPs were significantly associated with ESCC risk in the replication 2 samples in the same direction as in the genome-wide scan and replication 1 ($P = 2.20 \times 10^{-3}$ to $P = 1.67 \times 10^{-24}$). A joint analysis of the genome-wide scan data together with the samples from replications 1 and 2 showed that these 15 associations reached genome-wide significance (all $P \leq 2.49 \times 10^{-11}$; **Tables 2** and **3**).

Eight of the significant makers were located at chromosome 4q23, which harbors a cluster of seven alcohol dehydrogenase superfamily genes (*ADH* genes). The top marker in this region was rs1042026 (odds ratio (OR) = 1.35, 95% CI 1.29–1.41, $P_{combined} = 1.78 \times 10^{-39}$), and the other seven SNPs in this region are in moderate linkage disequilibrium (LD) with rs1042026 ($r^2 = 0.30–0.49$), all of which provided significant marginal associations in the combined dataset ($P_{combined} = 1.26 \times 10^{-29}$ to $P_{combined} = 2.75 \times 10^{-20}$) (**Fig. 1a**). After conditioning on rs1042026, the association *P* values for the other seven SNPs increased by over 13 orders of magnitude, suggesting

Table 2 Nine SNPs with significant marginal effects only on ESCC risk in the genome-wide discovery, replication and combined samples

SNP; chromosome; location (bp); gene; risk allele	Genome-wide discovery 2,031 cases, 2,044 controls			Replication 1 3,571 cases, 3,602 controls			Replication 2 4,521 cases, 5,018 controls			Combined sample 10,123 cases, 10,664 controls		
	MAF	OR (95% CI)	*P*	MAF	OR (95% CI)	*P*	MAF	OR (95% CI)	*P*	MAF	OR (95% CI)	*P*
rs4785204; chr. 16; 48,661,235; *HEATR3*; T	0.25	1.30 (1.17–1.43)	3.05×10^{-7}	0.27	1.23 (1.14–1.33)	9.45×10^{-8}	0.25	1.22 (1.14–1.30)	3.07×10^{-8}	0.26	1.24 (1.18–1.29)	2.24×10^{-20}
rs7206735; chr. 16; 48,706,009; *HEATR3*; C	0.28	1.29 (1.17–1.42)	3.49×10^{-7}	0.29	1.21 (1.12–1.30)	6.97×10^{-7}	0.28	1.18 (1.10–1.26)	1.31×10^{-6}	0.28	1.20 (1.15–1.26)	1.97×10^{-16}
rs6503659; chr. 17; 37,150,790; *HAP1*; A	0.12	1.39 (1.22–1.58)	5.11×10^{-7}	0.15	1.20 (1.09–1.31)	1.00×10^{-4}	0.12	1.27 (1.16–1.39)	2.36×10^{-7}	0.13	1.27 (1.20–1.34)	2.73×10^{-16}
rs2239815; chr. 22; 27,522,670; *XBP1*; T	0.38	1.28 (1.16–1.40)	1.72×10^{-7}	0.39	1.12 (1.04–1.20)	2.10×10^{-3}	0.35	1.17 (1.10–1.25)	9.17×10^{-7}	0.37	1.18 (1.13–1.23)	3.88×10^{-15}
rs2239612; chr. 3; 188,275,936; *ST6GAL1*; T	0.17	1.35 (1.20–1.51)	3.27×10^{-7}	0.20	1.20 (1.11–1.30)	1.15×10^{-5}	0.18	1.17 (1.08–1.26)	1.00×10^{-4}	0.19	1.21 (1.15–1.27)	5.74×10^{-14}
rs17761864; chr. 17; 2,118,387; *SMG6*; A	0.14	1.38 (1.22–1.56)	2.16×10^{-7}	0.15	1.16 (1.06–1.27)	1.60×10^{-3}	0.13	1.16 (1.06–1.27)	9.00×10^{-4}	0.14	1.21 (1.14–1.28)	2.21×10^{-11}
rs2847281; chr. 18; 12,811,593; *PTPN2*; C	0.15	1.33 (1.19–1.50)	1.37×10^{-6}	0.17	1.16 (1.06–1.27)	9.00×10^{-4}	0.14	1.14 (1.05–1.24)	2.20×10^{-3}	0.16	1.20 (1.14–1.26)	2.49×10^{-11}
rs4822983[a]; chr. 22; 27,445,066; *CHEK2*; T	0.19	1.46 (1.31–1.62)	1.02×10^{-8}	0.21	1.22 (1.12–1.32)	1.82×10^{-6}	0.19	1.24 (1.15–1.34)	2.06×10^{-8}	0.20	1.27 (1.21–1.34)	1.94×10^{-22}
rs1033667[a]; chr. 22; 27,460,300; *CHEK2*; T	0.26	1.33 (1.20–1.46)	1.91×10^{-8}	0.27	1.17 (1.09–1.26)	3.72×10^{-5}	0.24	1.26 (1.18–1.36)	3.69×10^{-11}	0.25	1.25 (1.19–1.30)	4.85×10^{-22}

The *P* values shown are two sided and were calculated by the additive model in a logistic regression analysis with age, sex, smoking, drinking and the first three principal components (for the GWAS stage only) as covariates. [a]Discovered by imputation analysis. Chr., chromosome; MAF, minor allele frequency in controls; OR, odds ratio for the minor allele.

Table 3 Eight SNPs with a significant marginal effect on ESCC risk and the interaction of genes and alcohol drinking in the genome-wide discovery, replication and combined samples

SNP; chromosome; location (bp); gene; substitution	Subgroup	Genome-wide discovery			Replication 1			Replication 2			Combined sample		
		MAF	OR (95% CI)	*P*	MAF	OR (95% CI)	*P*	MAF	OR (95% CI)	*P*	MAF	OR (95% CI)	*P*
rs1042026; chr. 4; 100,447,489; *ADH1B*; G>A	Case_control	0.26	1.29 (1.17–1.42)	1.51×10^{-7}	0.23	1.29 (1.19–1.39)	1.33×10^{-10}	0.22	1.44 (1.34–1.54)	1.67×10^{-24}	0.23	1.35 (1.29–1.41)	1.78×10^{-39}
	Nondrinker	0.24	1.11 (0.96–1.28)	1.63×10^{-1}	0.23	1.15 (1.04–1.28)	6.80×10^{-3}	0.21	1.31 (1.18–1.45)	4.91×10^{-7}	0.22	1.20 (1.12–1.27)	4.21×10^{-8}
	Drinker	0.27	1.44 (1.26–1.63)	3.69×10^{-8}	0.24	1.45 (1.29–1.63)	3.05×10^{-10}	0.22	1.58 (1.43–1.75)	1.94×10^{-19}	0.24	1.51 (1.42–1.61)	2.34×10^{-36}
	G × E		1.31 (1.08–1.59)	5.20×10^{-3}		1.28 (1.09–1.49)	2.00×10^{-3}		1.20 (1.04–1.38)	1.12×10^{-2}		1.27 (1.16–1.38)	2.54×10^{-7}
rs3805322; chr. 4; 100,276,021; *ADH4*;A>G	Case_control	0.48	0.79 (0.73–0.86)	1.89×10^{-7}	0.49	0.80 (0.75–0.86)	2.83×10^{-10}	0.48	0.84 (0.79–0.89)	1.25×10^{-8}	0.48	0.81 (0.78–0.85)	2.92×10^{-24}
	Nondrinker	0.49	0.94 (0.82–1.06)	2.93×10^{-1}	0.49	0.85 (0.77–0.93)	4.00×10^{-4}	0.48	0.95 (0.86–1.04)	2.35×10^{-1}	0.49	0.90 (0.85–0.95)	2.00×10^{-4}
	Drinker	0.47	0.68 (0.60–0.77)	1.23×10^{-9}	0.47	0.76 (0.68–0.84)	4.12×10^{-7}	0.47	0.76 (0.69–0.83)	1.47×10^{-9}	0.47	0.74 (0.70–0.78)	4.28×10^{-24}
	G × E		0.72 (0.60–0.85)	2.00×10^{-4}		0.88 (0.76–1.01)	7.48×10^{-2}		0.81 (0.71–0.91)	7.00×10^{-4}		0.81 (0.75–0.88)	5.58×10^{-7}
rs17033; chr. 4; 100,447,968; *ADH1B*; A>G	Case_control	0.11	1.40 (1.23–1.59)	2.80×10^{-7}	0.12	1.36 (1.23–1.50)	1.83×10^{-9}	0.10	1.44 (1.31–1.58)	1.98×10^{-14}	0.11	1.41 (1.33–1.50)	1.26×10^{-29}
	Nondrinker	0.11	1.31 (1.08–1.58)	6.50×10^{-3}	0.11	1.27 (1.11–1.45)	6.00×10^{-4}	0.10	1.35 (1.18–1.55)	1.29×10^{-5}	0.11	1.31 (1.21–1.43)	1.88×10^{-10}
	Drinker	0.12	1.47 (1.24–1.76)	1.69×10^{-5}	0.12	1.48 (1.27–1.72)	4.43×10^{-7}	0.10	1.50 (1.31–1.72)	4.02×10^{-9}	0.11	1.51 (1.38–1.64)	2.10×10^{-20}
	G × E		1.13 (0.87–1.46)	3.70×10^{-1}		1.18 (0.97–1.45)	1.03×10^{-1}		1.11 (0.92–1.33)	2.87×10^{-1}		1.14 (1.01–1.28)	3.23×10^{-2}
rs17028973; chr. 4; 100,541,809; *ADH7*; C>T	Case_control	0.34	1.26 (1.15–1.38)	4.61×10^{-7}	0.33	1.22 (1.14–1.31)	2.48×10^{-8}	0.31	1.29 (1.21–1.38)	7.67×10^{-15}	0.32	1.26 (1.21–1.32)	2.53×10^{-28}
	Nondrinker	0.32	1.16 (1.01–1.33)	3.05×10^{-2}	0.33	1.08 (0.98–1.19)	1.02×10^{-1}	0.31	1.17 (1.07–1.29)	1.10×10^{-3}	0.32	1.14 (1.07–1.21)	1.36×10^{-5}
	Drinker	0.36	1.35 (1.19–1.53)	2.29×10^{-6}	0.32	1.43 (1.28–1.59)	1.60×10^{-10}	0.31	1.42 (1.29–1.55)	2.58×10^{-13}	0.33	1.41 (1.33–1.50)	5.76×10^{-29}
	G × E		1.18 (0.99–1.42)	7.15×10^{-2}		1.34 (1.16–1.55)	7.20×10^{-5}		1.18 (1.04–1.34)	1.15×10^{-2}		1.24 (1.14–1.35)	3.04×10^{-7}
rs1614972; chr. 4; 100,477,178; *ADH1C*; T>C	Case_control	0.26	1.27 (1.15–1.40)	1.34×10^{-6}	0.25	1.26 (1.17–1.36)	1.62×10^{-9}	0.24	1.32 (1.23–1.41)	7.87×10^{-15}	0.25	1.28 (1.23–1.34)	8.02×10^{-28}
	Nondrinker	0.25	1.16 (1.00–1.33)	4.71×10^{-2}	0.25	1.16 (1.04–1.28)	5.20×10^{-3}	0.24	1.17 (1.06–1.29)	2.80×10^{-3}	0.25	1.16 (1.09–1.23)	6.18×10^{-6}
	Drinker	0.27	1.36 (1.19–1.56)	5.64×10^{-6}	0.25	1.40 (1.25–1.58)	1.85×10^{-8}	0.24	1.46 (1.32–1.62)	1.38×10^{-13}	0.25	1.42 (1.33–1.52)	2.39×10^{-26}
	G × E		1.20 (0.99–1.45)	7.06×10^{-2}		1.23 (1.06–1.44)	7.90×10^{-3}		1.25 (1.09–1.44)	1.60×10^{-3}		1.24 (1.13–1.36)	2.54×10^{-6}
rs1229977; chr. 4; 100,421,437; *ADH1A*; *C*>T	Case_control	0.10	1.37 (1.20–1.57)	3.22×10^{-6}	0.11	1.31 (1.19–1.45)	1.52×10^{-7}	0.11	1.33 (1.21–1.46)	4.41×10^{-9}	0.11	1.33 (1.25–1.41)	2.75×10^{-20}
	Nondrinker	0.10	1.19 (0.97–1.46)	9.89×10^{-2}	0.11	1.23 (1.07–1.41)	2.90×10^{-3}	0.10	1.25 (1.09–1.44)	1.80×10^{-3}	0.10	1.22 (1.12–1.33)	4.89×10^{-6}
	Drinker	0.11	1.49 (1.25–1.79)	1.28×10^{-5}	0.11	1.41 (1.21–1.65)	1.28×10^{-5}	0.11	1.43 (1.25–1.63)	2.00×10^{-7}	0.11	1.45 (1.33–1.58)	8.27×10^{-17}
	G × E		1.29 (0.99–1.70)	6.20×10^{-2}		1.17 (0.95–1.43)	1.46×10^{-1}		1.13 (0.93–1.36)	2.14×10^{-1}		1.19 (1.06–1.35)	4.30×10^{-3}
rs1789903; chr. 4; 100,481,064; *ADH1C*; C>G	Case_control	0.09	1.41 (1.22–1.62)	3.35×10^{-6}	0.09	1.31 (1.18–1.47)	1.04×10^{-6}	0.08	1.40 (1.26–1.55)	1.53×10^{-10}	0.09	1.37 (1.28–1.46)	9.46×10^{-21}
	Nondrinker	0.09	1.27 (1.03–1.58)	2.74×10^{-2}	0.10	1.13 (0.98–1.31)	9.21×10^{-2}	0.08	1.34 (1.15–1.56)	2.00×10^{-4}	0.09	1.24 (1.13–1.36)	6.67×10^{-6}
	Drinker	0.09	1.54 (1.26–1.88)	2.02×10^{-5}	0.09	1.55 (1.30–1.84)	6.37×10^{-7}	0.09	1.45 (1.25–1.68)	6.83×10^{-7}	0.09	1.50 (1.36–1.65)	1.16×10^{-16}
	G × E		1.25 (0.93–1.67)	1.34×10^{-1}		1.40 (1.12–1.75)	3.20×10^{-3}		1.08 (0.88–1.33)	4.64×10^{-1}		1.23 (1.08–1.41)	1.90×10^{-3}
rs1893883; chr. 4; 100,343,739; *ADH6*; C>G	Case_control	0.14	1.31 (1.16–1.48)	1.23×10^{-5}	0.14	1.30 (1.18–1.43)	3.61×10^{-8}	0.13	1.39 (1.28–1.52)	3.98×10^{-14}	0.14	1.34 (1.27–1.42)	2.69×10^{-25}
	Nondrinker	0.13	1.21 (1.01–1.45)	4.34×10^{-2}	0.14	1.21 (1.06–1.37)	3.40×10^{-3}	0.13	1.37 (1.21–1.56)	1.18×10^{-6}	0.13	1.26 (1.16–1.36)	8.48×10^{-9}
	Drinker	0.15	1.39 (1.18–1.63)	9.57×10^{-5}	0.15	1.41 (1.22–1.63)	2.04×10^{-6}	0.14	1.45 (1.29–1.64)	1.76×10^{-9}	0.15	1.42 (1.31–1.54)	2.84×10^{-18}
	G × E		1.17 (0.92–1.49)	2.02×10^{-1}		1.19 (0.99–1.44)	7.11×10^{-2}		1.03 (0.87–1.23)	7.09×10^{-1}		1.14 (1.02–1.27)	2.20×10^{-2}

The *P* values shown are two sided and were calculated by an additive model in a logistic regression analysis with age, sex, smoking, drinking and the first three principal components (for the GWAS stage only) as covariates for the subgroups of case_control (individuals included in the case-control study), nondrinker and drinker. The *P* values for the gene × environment interaction were calculated by conducting a 1-degree-of-freedom Wald test of a single interaction parameter (SNP × drinking status) as implemented in an unconditional logistic regression with age, sex, smoking as covariates. MAF, minor allele frequency in the controls; OR, odds ratio for the minor allele; G × E, gene × environment interaction.

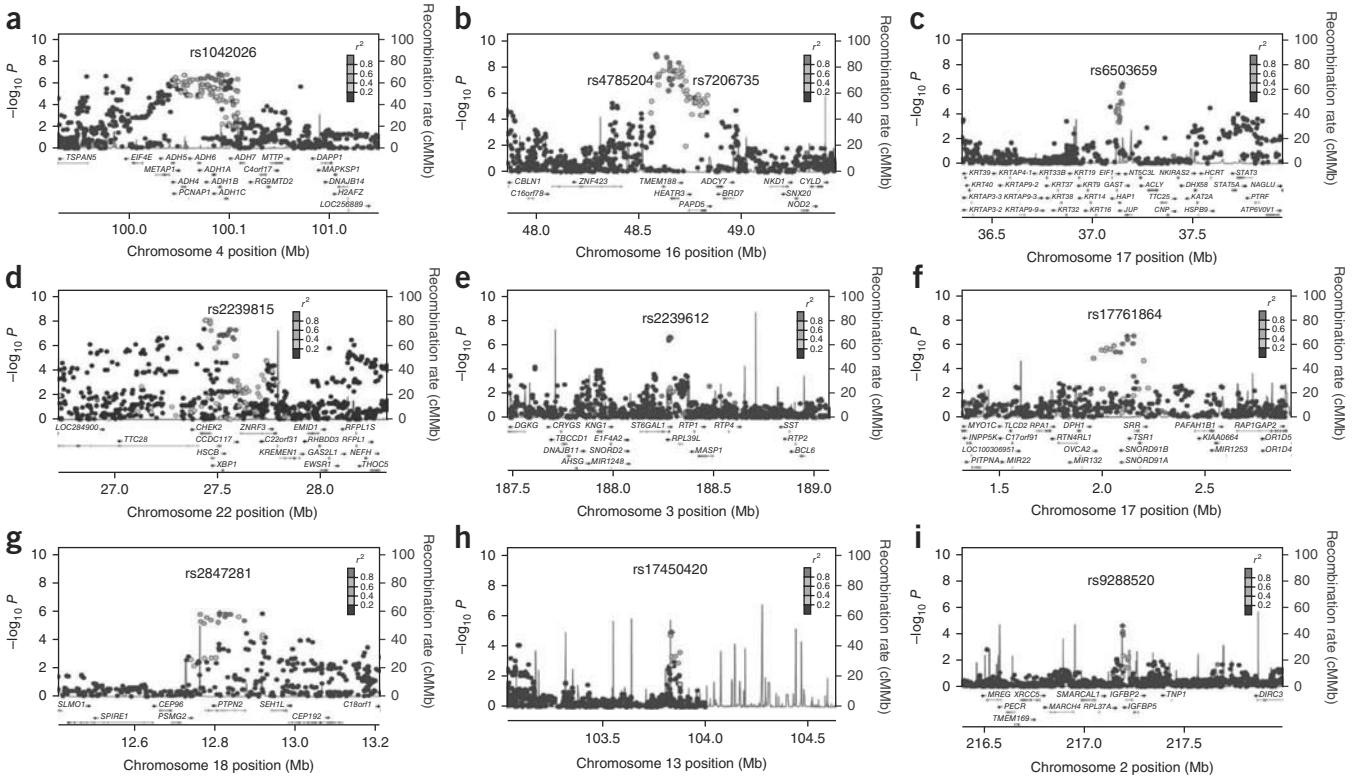


Figure 1 Regional plots of the association results for genotyped and imputed SNPs and the recombination rates within nine significant susceptibility loci. (**a**–**i**) The significant loci are located in chromosomes 4q23 (**a**),16q12.1 (**b**),17q21 (**c**), 22q12 (**d**), 3q27 (**e**), 17p13 (**f**), 18p11 (**g**), 13q33 (**h**) and 2q22 (**i**). For each plot, the $-\log_{10} P$ values (*y* axis) of the SNPs are shown according to their chromosomal positions (*x* axis). The estimated recombination rates (cM/Mb) from the HapMap Project (NCBI Build 36) are shown as light blue lines, and the genomic locations of genes within the regions of interest in the NCBI Build 36 human assembly were annotated from the UCSC Genome Browser and are shown as arrows. SNPs shown in red, orange, green, light blue and blue have $r^2 \geq 0.8$, $r^2 \geq 0.6$, $r^2 \geq 0.4$, $r^2 \geq 0.2$ and $r^2 < 0.2$ with the tag SNP, respectively. Purple diamonds represent associations of tag SNPs identified in the GWAS stage.

that the association signals of these other seven SNPs probably point toward the same locus, which is marked by the top SNP (rs1042026) (**Supplementary Table 2**). An imputation analysis in the initial GWAS identified associations for 111 SNPs within a 2-Mb region centered on rs1042026 ($P \leq 10^{-4}$), but none of these SNPs was more significant than the index marker, rs1042026, and a conditional analysis did not provide evidence for a second, independent susceptibility allele in this region (**Supplementary Table 3**).

The markers rs4785204 and rs7206735 at 16q12.1 were also strong signals that were associated with ESSC risk (OR = 1.24, 95% CI 1.18–1.29, $P_{combined} = 2.24 \times 10^{-20}$ for rs4785204 and OR = 1.20, 95% CI 1.15–1.26, $P_{combined} = 1.97 \times 10^{-16}$ for rs7206735; **Fig. 1b**). These two SNPs are located in close proximity to each other and are in moderate LD ($r^2 = 0.41$ in controls); after conditioning on rs4785204, rs7206735 was no longer genome-wide significant (**Supplementary Table 2**). An imputation analysis identified 40 untyped SNPs clustering in two blocks with high LD tagged by these markers that reached a significance of $P \leq 10^{-4}$; again, after conditioning on the index SNP (rs4785204), there was little evidence of a second susceptibility allele in this region (**Supplementary Table 3**).

We found new susceptibility loci for rs6503659 at 17q21 (OR = 1.27, 95% CI 1.20–1.34, $P_{combined} = 2.73 \times 10^{-16}$) and rs2239815 at 22q12 (OR = 1.18, 95% CI 1.13–1.23, $P_{combined} = 3.88 \times 10^{-15}$). Although we detected residual associations at many imputed SNPs in the region tagged by rs6503659, none of them was more significant than the index marker, and conditional analyses did not suggest the presence of a second susceptibility allele in this region (**Fig. 1c** and **Supplementary Table 3**). However, of the 36 imputed SNPs with $P \leq 10^{-4}$ in the region tagged by rs2239815, 8 comprised a separate significant block that was only in weak LD with rs2239815 ($r^2 = 0.21$–0.39) (**Fig. 1d**). We selected the top two imputed SNPs from this block, rs4822983 and rs1033667, both of which had *P* values that were smaller than that of the genotyped index SNP, rs2239815, in the initial GWAS for further replication in all samples. We found that each of these two SNPs was significantly associated with ESCC risk (OR = 1.27, 95% CI 1.21–1.34, $P_{combined} = 1.94 \times 10^{-22}$ for rs4822983 and OR = 1.25, 95% CI 1.19–1.30, $P_{combined} = 4.85 \times 10^{-22}$ for rs1033667; **Table 2**). After conditioning on rs4822983 in the combined sample, evidence for the associations between rs2239815 and rs1033667 and ESCC dropped by over ten orders of magnitude (**Supplementary Table 2**); similarly, after conditioning on rs4822983 in the initial GWAS, there was little evidence of a second susceptibility marker among the imputed SNPs in this region (**Supplementary Table 3**).

The SNP rs2239612 at 3q27 was also associated with ESCC risk (OR = 1.21, 95% CI 1.15–1.27, $P_{combined} = 5.74 \times 10^{-14}$), all seven imputed SNPs in this region were in strong LD with rs2239612 ($r^2 = 0.94$–0.99), and we identified no other significant LD block in this region (**Fig. 1e** and **Supplementary Table 3**). An additional two new identified markers were rs17761864 at 17p13 (OR = 1.21, 95% CI 1.14–1.28, $P_{combined} = 2.21 \times 10^{-11}$) and rs2847281 at 18p11 (OR = 1.20, 95% CI 1.14–1.26, $P_{combined} = 2.49 \times 10^{-11}$). An imputation analysis identified

Table 4 Two SNPs significantly associated with ESCC risk revealed by a SNP × alcohol drinking interaction analysis in the genome-wide discovery, replication and combined samples

SNP; chromosome; location (bp); gene; substitution	Subgroup	Genome-wide discovery			Replication 1			Replication 2			Combined sample		
		MAF	OR (95% CI)	*P*	MAF	OR (95% CI)	*P*	MAF	OR (95% CI)	*P*	MAF	OR (95% CI)	*P*
rs9288520; chr. 2; 217,189,516; *IGFB2*; G>A	Nondrinker	0.37	0.69 (0.60–0.79)	6.54×10^{-8}	0.33	0.87 (0.79–0.95)	3.40×10^{-3}	0.34	0.82 (0.74–0.90)	2.60×10^{-5}	0.34	0.81 (0.77–0.86)	4.72×10^{-12}
	Drinker	0.31	1.18 (1.03–1.34)	1.77×10^{-2}	0.30	1.08 (0.97–1.21)	1.58×10^{-1}	0.32	1.06 (0.96–1.17)	2.34×10^{-1}	0.31	1.09 (1.02–1.16)	1.00×10^{-2}
	G × E	0.34	1.71 (1.41–2.06)	2.69×10^{-5}	0.32	1.24 (1.07–1.43)	3.70×10^{-3}	0.33	1.29 (1.13–1.47)	2.00×10^{-4}	0.33	1.33 (1.22–1.45)	4.39×10^{-11}
rs17450420; chr. 13; 103,837,147; *SLC10A2*; A>G	Nondrinker	0.05	0.62 (0.44–0.87)	6.00×10^{-3}	0.05	0.76 (0.61–0.93)	8.20×10^{-3}	0.04	0.89 (0.72–1.09)	2.59×10^{-1}	0.05	0.78 (0.68–0.89)	2.00×10^{-4}
	Drinker	0.03	1.74 (1.27–2.37)	6.00×10^{-4}	0.05	1.27 (0.99–1.62)	5.78×10^{-2}	0.04	1.22 (0.99–1.51)	6.73×10^{-2}	0.04	1.34 (1.16–1.54)	4.65×10^{-5}
	G × E	0.04	2.76 (1.75–4.37)	1.37×10^{-5}	0.05	1.68 (1.22–2.31)	1.50×10^{-3}	0.04	1.42 (1.07–1.90)	1.68×10^{-2}	0.05	1.70 (1.41–2.06)	4.80×10^{-8}

The *P* values shown are two sided and were calculated by an additive model in a logistic regression analysis with age, sex, smoking, drinking and the first three principal components (for the GWAS stage only) as covariates for the subgroups of nondrinker and drinker. *P* values for the gene × environment interaction were calculated by conducting a 1-degree-of-freedom Wald test of a single interaction parameter (SNP × drinking status) as implemented in an unconditional logistic regression with age, sex, smoking as covariates. MAF, minor allele frequency in the controls; OR, odds ratio for the minor allele; chr., chromosome; G × E, gene × environment interaction.

15 associated SNPs with weaker signals than those of rs17761864 and rs2847281 (**Fig. 1f**,**g** and **Supplementary Table 3**).

Risk lock found by genome-wide gene-environment analysis

We performed a genome-wide gene-environment interaction analysis using previous genome-wide–association scan data by testing whether the per-allele odds ratio for each SNP differed between ever drinkers and never drinkers. A quantile-quantile plot of the observed versus expected Wald χ^2 1-degree-of-freedom test for interaction showed no evidence for inflation ($\lambda = 1.004$; **Supplementary Fig. 1a**). There were 25 promising SNPs associated with ESCC risk at significance levels ranging from $P_{G \times E} = 1.42 \times 10^{-23}$ to $P_{G \times E} = 9.88 \times 10^{-5}$ (**Supplementary Fig. 1b**). Among them, 15 SNPs were located at 12q24, including rs11066015 ($P_{G \times E} = 1.42 \times 10^{-23}$), rs11066280 ($P_{G \times E} = 1.25 \times 10^{-17}$) and rs2074356 ($P_{G \times E} = 3.38 \times 10^{-16}$), which our previous report showed all interact with alcohol drinking to promote ESCC risk[12]. rs11066015 is in strong LD with rs671 ($r^2 = 0.79$), a functional SNP in *ALDH2* (encoding aldehyde dehydrogenase-2) that is known to be associated with both a flushing response to alcohol intake and ESCC risk in a drinking-behavior–specific manner[4,7–9]. We then performed a fast-track replication in the replication 1 samples of the ten remaining tag SNPs located in regions other than 12q24. Of these ten SNPs, eight did not replicate (all $P_{G \times E} > 0.05$; **Supplementary Table 4**), and we did not evaluate them further. Additional replication (replication 2) of rs9288520 at 2q22 and rs17450420 at 13q33 verified their associations with ESCC risk (combined-sample $P_{G \times E} = 4.39 \times 10^{-11}$ and $P_{G \times E} = 4.80 \times 10^{-8}$, respectively). The minor allele of rs9288520 was associated with reduced risk of ESCC in all nondrinkers (OR = 0.81, 95% CI 0.77–0.86, $P = 4.72 \times 10^{-12}$) but was associated with increased risk in all drinkers (OR = 1.09, 95% CI 1.02–1.16, $P = 0.01$). Similarly, the minor allele of rs17450420 was associated with reduced ESCC risk in nondrinkers (OR = 0.78, 95% CI 0.68–0.89, $P = 0.0002$) but with increased risk in drinkers (OR = 1.34, 95% CI 1.16–1.54, $P = 4.65 \times 10^{-5}$; **Table 4**). Neither of these two SNPs was significantly related to drinking status in cases, controls or the combined sample (data not shown). To increase the spectrum of variants tested, we performed an imputation analysis in the GWAS set and identified two and nine imputed SNPs, respectively, in the two regions that showed significant interactions with drinking ($P_{G \times E} < 10^{-4}$); however, none of these SNPs was more significantly related to cancer than the index markers in each region, rs17450420 and rs9288520 (**Fig. 1h**,**i**).

Alcohol use and interaction in ESCC susceptibility

Stratified analyses showed that the nine newly associated SNPs at 22q12, 17q21, 17p13, 16q12.1, 3q27 and 18p11 were all significantly associated with ESCC risk in the same direction in both drinkers and nondrinkers; the associations did not differ significantly between subgroups categorized by alcohol-drinking status (**Supplementary Table 5**). The associations for the eight SNPs (rs1042026, rs3805322, rs17028973, rs1614972, rs17033, rs1229977, rs1789903 and rs1893883) at 4q23 differed by alcohol use, with higher risk in drinkers than in nondrinkers (interaction $P = 2.54 \times 10^{-7}$ to $P = 3.23 \times 10^{-2}$; **Table 3**), which is consistent with previously published epidemiologic data[4,7,9,15,17]. An analysis of the joint effects of drinking, rs1042026 in *ADH1B* and rs11066015 in *ALDH2* on risk of developing ESCC identified that the odds in drinkers carrying risk alleles at both *ADH1B* (GA or AA genotype) and *ALDH2* (GA or AA genotype) was approximately fourfold higher than that in drinkers carrying the nonrisk *ADH1B* G and *ALDH2* G alleles and was more than threefold higher than that in nondrinkers carrying the risk alleles. The effect sizes of

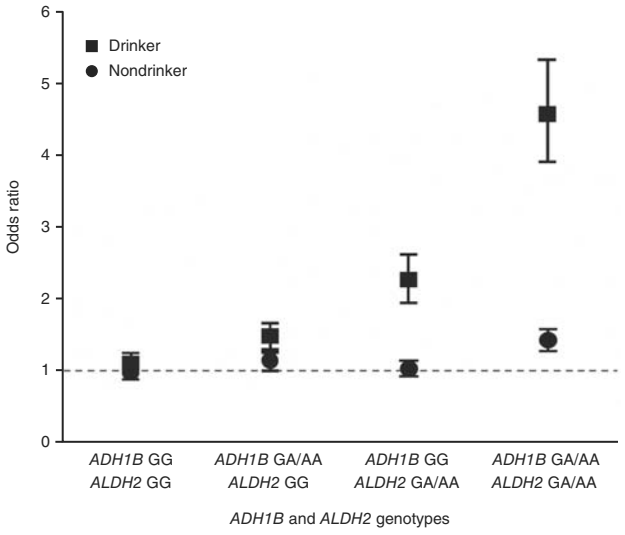


Figure 2 Plots showing the ORs for ESCC in alcohol drinkers and nondrinkers with different *ADH1B* rs1042026 and *ALDH2* rs11066015 genotypes. The vertical bars represent the 95% CIs. The horizontal dashed line indicates the null value (OR = 1.0).

the *ADH1B* and *ALDH2* variants for ESCC risk in nondrinkers were not large (**Fig. 2** and **Supplementary Table 6**).

Replication of susceptibility loci in a high-risk population

We next examined whether these significant loci were also associated with susceptibility to ESCC in 1,410 cases and 1,656 controls obtained from a high-risk population in Shanxi province, China, as described in a previous GWAS[11]. We found that among the 18 SNPs listed in **Tables 2–4**, 4 showed significant association (**Supplementary Table 7**) in this independent dataset. rs2239815 (OR = 1.24, 95% CI 1.12–1.38, $P = 3.23 \times 10^{-5}$), rs4822983 (OR = 1.28, 95% CI 1.07–1.54, $P = 0.0082$) and rs1033667 (OR = 1.35, 95% CI 1.20–1.51, $P = 2.95 \times 10^{-7}$) at 22q12 and rs2239612 (OR = 1.15, 95% CI 1.01–1.30, $P = 0.0343$) at 3q27 were all associated with increased ESCC risk in this group, as was observed in the other groups. A gene-drinking interaction analysis showed that rs1614972 in *ADH1C* at 4q23 had evidence for replication (OR = 1.37, 95% CI 1.03–1.82, $P_{G \times E} = 0.0281$). In this case-control group, there was some evidence for an interaction between rs2847281 at 18p11 (*PTPN2*) and alcohol drinking (OR = 1.55, 95% CI 1.12–2.15, $P_{G \times E} = 0.0083$).

DISCUSSION

In a multistage GWAS, we identified nine new susceptibility loci associated with ESCC risk across three independent study groups comprising a total of 10,123 cases and 10,664 controls. Among these loci, three had a significant interaction with alcohol drinking, an important lifestyle risk factor in the development of ESCC. We also confirmed some of our findings in an independent study from a high-risk population. To the best of our knowledge, this is one of the largest studies to explore gene-environment interactions for risk of developing ESCC by incorporating alcohol-drinking status into the primary GWAS stage 1 analysis.

Among the six regions with a notable marginal effect for risk of ESCC, two at 16q12.1 tagged by rs4785204 and rs7206735 contain the *TMEM188*, *HEATR3* and *PAPD5* genes, which are interesting and plausible candidate genes worthy of follow-up studies. The variant rs6503659 is located 13,595 bp downstream of *JUP* and 6,366 bp upstream of *HAP1* at 17q21. *JUP* encodes γ-catenin, a cytoplasmic protein that has a similar structure and function to β-catenin and serves as a cell-to-cell attachment molecule through its interaction with E-cadherin[18,19]. The role of γ-catenin in cancer is complex and dependent on the cellular context. Functional loss of γ-catenin results in tumor invasion or metastasis, and γ-catenin is an important part of Wnt signaling[20–22]. Therefore, subtle changes in γ-catenin expression caused by genetic variation could potentially influence the differentiation or invasion of certain transformed cells, resulting in cancer formation. *HAP1* produces huntingtin-associated protein-1, a binding partner of the Huntington's disease protein huntingtin. *HAP1* is involved in vesicular transport, gene transcription regulation, membrane receptor trafficking and other functions such as calcium release and protein aggregation[23,24]. However, the function of *HAP1* in cancer is not clear.

In a further imputation analysis, we identified two LD blocks at 22q12 that contain *XBP1* and *CHEK2*, which encode X-box binding protein 1 and a cell-cycle checkpoint kinase, respectively. *XBP1* is an important part of the unfolded protein response that is involved in the regulation of endoplasmic reticulum stress–mediated apoptosis, and aberrant expression of *XBP1* has been implicated in cancer development and progression, as well as in resistance to drugs[25–28]. *CHEK2* is responsible for preventing DNA-damaged cells from entering into mitosis, a crucial step to avert cancer development. It has therefore been considered as a candidate cancer susceptibility gene[29]. Markers near *CHEK2* have been found previously to be promising signals in the NCI GWAS[11] that we used for replication in this study. In view of the probable important roles of *XBP1* and *CHEK2* in cancer, it is plausible that genetic variations influencing the functions of these genes may confer susceptibility to ESCC.

rs2239612 is located at 3q27 in *ST6GAL1*, which encodes ST6 β-galactosamide α-2,6-sialyltransferase. Previous studies have shown that *ST6GAL1* is upregulated in many types of human cancers, and elevated expression of *ST6GAL1* is also correlated with tumor invasiveness and metastasis[30–33]. rs17761864 is located at 17p13 in *SMG6* (also known as *EST1A*), whose product is an essential factor in nonsense-mediated mRNA decay and telomere maintenance[34,35]; however, whether this gene has a role in cancer is currently unknown. The variant rs2847281 is located at 18p11 in *PTPN2*, encoding non-receptor type 2 protein tyrosine phosphatase, which not only influences the development of the immune system but is also linked to a number of autoimmune diseases and cancer[36,37].

In this study, we performed gene-environment interaction analyses by testing for differences in the per-allele odds ratios between ever drinkers and never drinkers. These analyses identified three genomic regions that had significant interactions with alcohol consumption to promote risk of developing ESCC. Notably, on chromosome 4q23 there is a region that harbors a cluster of seven genes encoding alcohol dehydrogenase (ADH) family proteins ((listed 5′ to 3′) *ADH7*, *ADH1C*, *ADH1B*, *ADH1A*, *ADH6*, *ADH4* and *ADH5*). ADHs oxidize alcohol to acetaldehyde, a carcinogen that is probably important in the etiology of alcohol-related cancers[38]. Drinkers with the fast ADH metabolizer genotype produce more acetaldehyde and are expected to have an elevated risk of these cancers. However, in this study, we were unable to determine the exact contribution of individual variants because of the LD pattern over the region covering the *ADH* genes. Deep sequencing of this region is warranted to map candidate genes and variants for follow-up functional analyses. Using a genome-wide gene-environment interaction analysis, we found that the most significant interaction region was for variants at 12q24 harboring *ALDH2*, which encodes aldehyde dehydrogenase-2 that, in turn, detoxifies acetaldehyde to acetate. The directions of our associations reported here are consistent with those reported in our previous GWAS[12] and other published studies[4,8,9]. Furthermore, in the present study, we evaluated the joint effects of *ADH1B* and *ALDH2* variants and drinking on ESCC risk and found that individuals who carried both of the risk alleles of *ADH1B* and *ALDH2* and were classified as alcohol drinkers had the highest risk. These findings clearly indicate a gene-environment interaction between alcohol use and genetic variation in the alcohol-metabolizing pathway for developing ESCC. Because ADHs oxidize alcohol to carcinogenic acetaldehyde, which is then detoxified by aldehyde dehydrogenases, it is anticipated that individuals with the combination of the fast alcohol metabolizer genotype and the slow acetaldehyde metabolizer genotype would be most susceptible to ESCC. These results strongly highlight the potential importance of reducing alcohol use in individuals carrying high-risk alleles to reduce ESCC risk.

We also identified associations with ESCC risk for rs9288520, located upstream of *IGFBP2* at 2q22, and rs17450420, located in a gene desert upstream of *SLC10A2* at 13q33. These two variants did not show marginal effects but were significantly associated with risk when alcohol drinking was incorporated into the genome-wide gene-environment interaction analysis. Compared to common alleles, the minor alleles of these two SNPs were associated with decreased risk of ESCC in nondrinkers and increased risk in drinkers. *IGFBP2* produces

insulin-like growth factor binding protein 2, which is involved in cell proliferation, migration and apoptosis, and elevated serum IGFBP2 concentrations have been detected in patients with various types of cancer[39]. Interestingly, it has been shown that *IGFBP2* RNA is overexpressed in the placenta and fetal lungs of rats fed with alcohol, and this overexpression is associated with ethanol-induced growth retardation[40]. *SLC10A2* encodes a sodium/bile acid cotransporter and has been suggested to be associated with alcohol dependence[41].

Because of the stringent *P* values we required for statistical significance to prevent false-positive findings in the GWAS, additional associations with promising *P* values were not confirmed in the present study, underscoring the need to continue the search for new loci[13]. Therefore, it is important to undertake complementary strategies to discover additional variants, particularly when some genetic effects are dependent on environmental exposure and may show a substantial effect only when a specific environmental exposure is present. Indeed, by replication of more potential associated SNPs in expanded samples and by performing a genome-wide gene-environment interaction analysis, we extended our GWAS results with the discovery of nine new ESCC susceptibility loci.

We also replicated the results in an additional case-control group from Shanxi province, a region with extremely high rates of ESCC in China[11]. We verified that four of the nine loci identified in the GWAS and replication samples also had significant marginal genetic effects on ESCC risk in this high-risk population; however, we found only modest evidence for a gene–alcohol drinking interaction in this population. This apparent inconsistency probably reflects differences in environmental exposures between general and high-risk populations. It is well known that in general populations, alcohol drinking and tobacco smoking are the major risk factors for ESCC[15,17]. However, in some high-risk regions of the Shanxi and Henan provinces in China, alcohol drinking has little or no association with ESCC risk[42,43], whereas other factors such as nutritional deficiencies, family history and certain chemical carcinogens in the diet are strongly associated with this type of cancer[44–46]. In this context, it is therefore not surprising to observe different genetic risks between general and high-risk populations. These differences also emphasize the importance of further analyses of interactions between genetic variants and the specific environmental factors in high-risk populations.

In conclusion, we have identified nine new susceptibility loci for ESCC in Chinese populations, extending our previous findings and advancing the understanding of the genetic etiology of ESCC. The newly identified susceptibility loci warrant follow-up fine-mapping and functional studies. Furthermore, the risk variants in the alcohol metabolism pathway that we have confirmed in this large study might be useful for identifying high-risk individuals for the prevention of ESCC in the Chinese population, particularly where alcohol consumption is a possible health risk.

URLs. R, http://www.r-project.org/; MACH, http://www.sph.umich.edu/csg/abecasis/MACH/index.html; LocusZoom, http://csg.sph.umich.edu/locuszoom/.

METHODS

Methods and any associated references are available in the online version of the paper.

Note: Supplementary information is available in the online version of the paper.

ACKNOWLEDGMENTS

This work was funded by the National High-Tech Research and Development Program of China (2009AA022706 to D.L.), the National Basic Research Program of China (2011CB504303 to D.L. and W.T.), the National Natural Science Foundation of China (30721001 to D.L., Q.Z. and Z.L.) and the Intramural Research Program of the US National Institutes of Health, NCI and the Division of Cancer Epidemiology and Genetics.

AUTHOR CONTRIBUTIONS

D.L. was the overall principle investigator of the study who conceived the study and obtained financial support, was responsible for study design, oversaw the entire study, interpreted the results and wrote parts of and synthesized the paper. C.W. performed overall project management, oversaw laboratory analyses, performed statistical analyses and drafted the initial manuscript. P.K. oversaw statistical analyses, interpreted the results and reviewed the manuscript. Y. Li and L.L. performed the imputation analysis and reviewed the manuscript. K.Z., J.C., Y.Q., Yuling Zhou and Y. Liu performed laboratory analyses. Z. Hu, G.J. and H.S. were responsible for subject recruitment and sample preparation of Nanjing samples. Z. He, C.G., C.L., H.Y. and Y.K. were responsible for subject recruitment and sample preparation of Henan samples. W.J., J.F. and Y. Zeng were responsible for subject recruitment and sample preparation of Guangzhou samples. X.M. and T.W. provided some of the control samples. Yifeng Zhou was responsible for subject recruitment of the additional validation cohorts. D.Y. and W.T. performed subject recruitment and sample preparation of Beijing samples. Q.Z. and Z.L. provided some of the financial support and reviewed the manuscript. Z.W., C.C.A., N.H., N.D.F., T.D., A.M.G., S.J.C. and P.R.T. performed subject recruitment, sample preparation, laboratory analysis and statistical analysis of Shanxi samples and reviewed the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Published online at http://www.nature.com/doifinder/10.1038/ng.2411.
Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Sun, T. *et al.* Polymorphisms of death pathway genes *FAS* and *FASL* in esophageal squamous-cell carcinoma. *J. Natl. Cancer Inst.* **96**, 1030–1036 (2004).
2. Zhang, X. *et al.* Identification of functional genetic variants in *cyclooxygenase-2* and their association with risk of esophageal cancer. *Gastroenterology* **129**, 565–576 (2005).
3. Sun, T. *et al.* A six-nucleotide insertion-deletion polymorphism in the *CASP8* promoter is associated with susceptibility to multiple cancers. *Nat. Genet.* **39**, 605–613 (2007).
4. Lewis, S.J. & Smith, G.D. Alcohol, ALDH2, and esophageal cancer: a meta-analysis which illustrates the potentials and limitations of a Mendelian randomization approach. *Cancer Epidemiol. Biomarkers Prev.* **14**, 1967–1971 (2005).
5. Hiyama, T., Yoshihara, M., Tanaka, S. & Chayama, K. Genetic polymorphisms and esophageal cancer risk. *Int. J. Cancer* **121**, 1643–1658 (2007).
6. Akbari, M.R. *et al.* Candidate gene association study of esophageal squamous cell carcinoma in a high-risk region in Iran. *Cancer Res.* **69**, 7994–8000 (2009).
7. Hashibe, M. *et al.* Multiple *ADH* genes are associated with upper aerodigestive tract cancers. *Nat. Genet.* **40**, 707–709 (2008).
8. McKay, J.D. *et al.* A genome-wide association study of upper aerodigestive tract cancers conducted within the INHANCE consortium. *PLoS Genet.* **7**, e1001333 (2011).
9. Cui, R. *et al.* Functional variants in *ADH1B* and *ALDH2* coupled with alcohol and smoking synergistically enhance esophageal cancer risk. *Gastroenterology* **137**, 1768–1775 (2009).
10. Wang, L.D. *et al.* Genome-wide association study of esophageal squamous cell carcinoma in Chinese subjects identifies susceptibility loci at *PLCE1* and *C20orf54*. *Nat. Genet.* **42**, 759–763 (2010).
11. Abnet, C.C. *et al.* A shared susceptibility locus in *PLCE1* at 10q23 for gastric adenocarcinoma and esophageal squamous cell carcinoma. *Nat. Genet.* **42**, 764–767 (2010).
12. Wu, C. *et al.* Genome-wide association study identifies three new susceptibility loci for esophageal squamous-cell carcinoma in Chinese populations. *Nat. Genet.* **43**, 679–684 (2011).
13. Panagiotou, O.A. & Ioannidis, J.P. What should the genome-wide significance threshold be? Empirical replication of borderline genetic associations. *Int. J. Epidemiol.* **41**, 273–286 (2012).
14. Park, J.H. *et al.* Estimation of effect size distribution from genome-wide association studies and implications for future discoveries. *Nat. Genet.* **42**, 570–575 (2010).
15. Islami, F. *et al.* Alcohol drinking and esophageal squamous cell carcinoma with focus on light-drinkers and never-smokers: a systematic review and meta-analysis. *Int. J. Cancer* **129**, 2473–2484 (2011).
16. Hunter, D.J. Gene-environment interactions in human diseases. *Nat. Rev. Genet.* **6**, 287–298 (2005).
17. Gao, Y.T. *et al.* Risk factors for esophageal cancer in Shanghai, China. I. Role of cigarette smoking and alcohol drinking. *Int. J. Cancer* **58**, 192–196 (1994).

18. Aberle, H., Schwartz, H. & Kemler, R. Cadherin-catenin complex: protein interactions and their implications for cadherin function. *J. Cell Biochem.* **61**, 514–523 (1996).
19. Bullions, L.C. & Levine, A.J. The role of β-catenin in cell adhesion, signal transduction, and cancer. *Curr. Opin. Oncol.* **10**, 81–87 (1998).
20. Morin, P.J. *et al.* Activation of β-catenin–Tcf signaling in colon cancer by mutations in β-catenin or APC. *Science* **275**, 1787–1790 (1997).
21. Kolligs, F.T. *et al.* γ-catenin is regulated by the APC tumor suppressor and its oncogenic activity is distinct from that of β-catenin. *Genes Dev.* **14**, 1319–1331 (2000).
22. Simcha, I. *et al.* Suppression of tumorigenicity by plakoglobin: an augmenting effect of N-cadherin. *J. Cell Biol.* **133**, 199–209 (1996).
23. Li, X.J. *et al.* A huntingtin-associated protein enriched in brain with implications for pathology. *Nature* **378**, 398–402 (1995).
24. Wu, L.L. & Zhou, X.F. Huntingtin associated protein 1 and its functions. *Cell Adh. Migr.* **3**, 71–76 (2009).
25. Kim, R., Emi, M., Tanabe, K. & Murakami, S. Role of the unfolded protein response in cell death. *Apoptosis* **11**, 5–13 (2006).
26. Koong, A.C., Chauhan, V. & Romero-Ramirez, L. Targeting XBP-1 as a novel anti-cancer strategy. *Cancer Biol. Ther.* **5**, 756–759 (2006).
27. Romero-Ramirez, L. *et al.* XBP1 is essential for survival under hypoxic conditions and is required for tumor growth. *Cancer Res.* **64**, 5943–5947 (2004).
28. Shuda, M. *et al.* Activation of the *ATF6*, *XBP1* and *grp78* genes in human hepatocellular carcinoma: a possible involvement of the ER stress pathway in hepatocarcinogenesis. *J. Hepatol.* **38**, 605–614 (2003).
29. Antoni, L., Sodha, N., Collins, I. & Garrett, M.D. CHK2 kinase: cancer susceptibility and cancer therapy—two sides of the same coin? *Nat. Rev. Cancer* **7**, 925–936 (2007).
30. Dall'Olio, F., Chiricolo, M. & Lau, J.T. Differential expression of the hepatic transcript of β-galactoside α-2,6-sialyltransferase in human colon cancer cell lines. *Int. J. Cancer* **81**, 243–247 (1999).
31. Wang, P.H. *et al.* Enhanced expression of α 2,6-sialyltransferase ST6Gal I in cervical squamous cell carcinoma. *Gynecol. Oncol.* **89**, 395–401 (2003).
32. Recchi, M.A. *et al.* Multiplex reverse transcription polymerase chain reaction assessment of sialyltransferase expression in human breast cancer. *Cancer Res.* **58**, 4066–4070 (1998).
33. Pousset, D., Piller, V., Bureaud, N., Monsigny, M. & Piller, F. Increased α 2,6 sialylation of N-glycans in a transgenic mouse model of hepatocellular carcinoma. *Cancer Res.* **57**, 4249–4256 (1997).
34. Eberle, A.B., Lykke-Andersen, S., Mühlemann, O. & Jensen, T.H. SMG6 promotes endonucleolytic cleavage of nonsense mRNA in human cells. *Nat. Struct. Mol. Biol.* **16**, 49–55 (2009).
35. DeZwaan, D.C. & Freeman, B.C. The conserved Est1 protein stimulates telomerase DNA extension activity. *Proc. Natl. Acad. Sci. USA* **106**, 17337–17342 (2009).
36. Doody, K.M., Bourdeau, A. & Tremblay, M.L. T-cell protein tyrosine phosphatase is a key regulator in immune cell signaling: lessons from the knockout mouse model and implications in human disease. *Immunol. Rev.* **228**, 325–341 (2009).
37. Dubé, N. & Tremblay, M.L. Involvement of the small protein tyrosine phosphatases TC-PTP and PTP1B in signal transduction and diseases: from diabetes, obesity to cell cycle, and cancer. *Biochim. Biophys. Acta* **1754**, 108–117 (2005).
38. World Cancer Research Fund/American Institute for Cancer Research. *Food, Nutrition, Physical Activity, and the Prevention of Cancer: a Global Perspective* (AICR, Washington, DC, 2007).
39. Hoeflich, A. *et al.* Insulin-like growth factor-binding protein 2 in tumorigenesis: protector or promoter? *Cancer Res.* **61**, 8601–8610 (2001).
40. Fatayerji, N., Engelmann, G.L., Myers, T. & Handa, R.J. *In utero* exposure to ethanol alters mRNA for insulin-like growth factors and insulin-like growth factor-binding proteins in placenta and lung of fetal rats. *Alcohol. Clin. Exp. Res.* **20**, 94–100 (1996).
41. Edenberg, H.J. *et al.* Genome-wide association study of alcohol dependence implicates a region on chromosome 11. *Alcohol. Clin. Exp. Res.* **34**, 840–852 (2010).
42. Gao, Y. *et al.* Risk factors for esophageal and gastric cancers in Shanxi Province, China: a case-control study. *Cancer Epidemiol.* **35**, e91–e99 (2011).
43. He, Z. *et al.* Prevalence and risk factors for esophageal squamous cell cancer and precursor lesions in Anyang, China: a population-based endoscopic survey. *Br. J. Cancer* **103**, 1085–1088 (2010).
44. Mark, S.D. *et al.* Prospective study of serum selenium levels and incident esophageal and gastric cancers. *J. Natl. Cancer Inst.* **92**, 1753–1763 (2000).
45. Gao, Y. *et al.* Family history of cancer and risk for esophageal and gastric cancer in Shanxi, China. *BMC Cancer* **9**, 269 (2009).
46. Lu, S.H. *et al.* Relevance of N-nitrosamines to esophageal cancer in China. *J. Cell Physiol. Suppl.* **4**, 51–58 (1986).

ONLINE METHODS

Study subjects. This study was an extension of our previous GWAS in which the genome-wide scan sample comprised 2,031 cases with ESCC and 2,044 controls and the replication samples comprised 6,276 cases and 6,165 controls. The sources and characteristics of these study subjects were described previously[12]. To further increase our statistical power for validation, we added an additional 1,816 cases and 2,455 controls in the present study. These cases and controls were recently recruited from the Han Chinese population through collaboration with multiple hospitals in Beijing and Jiangsu province, China. A diagnosis of ESCC was confirmed by either histopathologic or cytologic analyses, as described previously[12]. Demographic characteristics of the subjects, including age, sex, smoking status and drinking status, were obtained from each patient's medical records. Control subjects were selected on the basis of a physical examination and were frequency matched for age and sex to the cases with ESCC, as previously described[12]. All the cases and controls for each of the replication cohorts were sampled from the same locality and the same population to assure minimal population stratification. In replication 1, a total of 3,571 cases and 3,602 controls were collected from the Beijing region, and in replication 2, 4,521 cases and 5,018 controls were recruited from the Jiangsu, Henan and Guangdong provinces. This study also included an additional validation cohort consisting of 1,410 cases with ESCC and 1,656 controls from a study conducted in a population at high risk for ESCC in Shanxi, China, as described previously[11]. For this study, alcohol drinking status was assessed by a detailed questionnaire[42]. For the present analysis, individuals were classified as drinkers if they reported drinking any form of alcohol at least twice a week; otherwise, they were defined as nondrinkers. Individuals who reported smoking more than 100 cigarettes in their life or smoking tobacco in a pipe more than 100 times were defined as smokers; all others were defined as nonsmokers. The distributions of the selected characteristics among the cases and controls for each of the study sets examined in the genome-wide scan and in each replication are shown in **Table 1**. At recruitment, informed consent was obtained from each subject, and the study was approved by the institutional review boards of the Chinese Academy of Medical Sciences Cancer Institute, Peking University, SunYat-Sen University Cancer Center, Nanjing Medical University, the Medical College of Soochow University, Shanxi Cancer Hospital and the US NCI.

SNP selection and genotyping for replication. In replication 1, we selected SNPs with marginal significance ($10^{-7} < P \leq 10^{-4}$) for the genetic association analysis and SNPs with $P \leq 10^{-4}$ for the genome-wide gene × drinking interaction analysis. All selection was based on our previous GWAS scan results[12]. We adopted a two-step approach to select these SNPs. First, we excluded those SNPs with MAF < 0.01 in both cases and controls and those with genotype frequencies not conforming to Hardy-Weinberg equilibrium (HWE) in the controls ($P < 0.01$). Second, we computed the correlation coefficient (r) of each pair of adjacent SNPs on the same chromosome to assess LD status. SNPs with $r^2 > 0.8$ were considered to be in one LD block, and we thus selected the most significant SNP (with the lowest P value) in the block for replication. Using these criteria, we selected 175 SNPs for the genetic association analysis and 12 SNPs for the genome-wide gene × drinking interaction analysis in replication 1. Genotyping in replication cohort 1 was accomplished with an Illumina GoldenGate Assay of 187 attempted SNPs (Illumina). We filtered out SNPs with call rate <95% or with genotype frequencies in controls departing from HWE ($P < 0.01$). Finally, 169 and 10 genotyped SNPs passed quality control and were included in the final genetic association analysis and the final gene × drinking interaction analysis, respectively. We next selected SNPs with association at a significance of $P < 0.01$ for the replication 2 analysis. With this criterion, 18 SNPs for the genetic association analysis and 2 SNPs for the gene × drinking interaction analysis were selected and genotyped using a TaqMan genotyping platform (ABI 7900HT Real Time PCR system, Applied Biosystems) in replication 2.

For genotyping quality control, we implemented several measures in the replication assays, including (i) case and control samples were mixed in the plates, (ii) persons who performed the genotyping assays were not aware of the case or control status of the samples, (iii) both positive and negative (no DNA) control samples were included on every 384-well assay plate and (iv) replication of nearly 10% of the total DNA samples (400 in the GWAS scan and 700 in replication 1) was performed using the TaqMan genotyping platform (with duplication concordances of 99.92% and 99.99%, respectively).

Statistical analyses. For the GWAS, associations between genotypes and risk of developing ESCC were analyzed by an additive model in a logistic regression (genotypic trend effect with a 1-degree-of-freedom test) framework with age, sex, smoking, drinking and the first three principal components from EIGENSTRAT as covariates[12]. SNPs imputed using the GWAS scan data were included in this logistic regression model using SNP 'dosages' (the expected allele counts). Conditional association analyses were conducted by including in the unconditional logistic regression model the most significant SNP on 4q23, 16q12.1 or 22q12 and examining the association between each of the remaining SNPs and risk of ESCC. For the analysis of the gene × drinking interaction, we tested the interaction between each SNP and drinking status by conducting a 1-degree-of-freedom Wald test of a single interaction parameter (SNP × drinking status) as implemented in an unconditional logistic regression based on the equation $Y = \beta_0 + \beta_1 \times \text{SNP} + \beta_2 \times \text{drinking status} + \beta_3 \times (\text{SNP} \times \text{drinking status})$. Here, Y is the logit of the probability of being a case, β_0 is a constant, β_1 and β_2 are the main effects of SNP and dinking status, respectively, and β_3 is the interaction term to be tested. We further performed stratified analyses of significant SNPs identified by a two-phase replication strategy in different cohorts: we used case or control status as the outcome and tested the associations in the nondrinker and drinker groups. Sex, age, smoking and first three principal components served as covariates in both the genome-wide gene × drinking interaction analysis and the stratified analysis. The odds ratios calculated are presented for the minor allele of each SNP. For fine mapping of the significant regions, we used MACH software (see URLs) to impute untyped markers using LD and haplotype information from the HapMap II CHB + JPT populations as the reference set. To identify susceptibility genes underlying the various associations, we analyzed the LD patterns around the risk-associated SNPs and determined LD blocks where the risk-associated SNPs were located. We then investigated the gene or genes covered by the LD blocks. The LD structures and haplotype block plots were generated using Haploview v4.1 software (see URLs). Significant regions were plotted using the online tool LocusZoom (see URLs).

doi:10.1038/ng.2411

Joint Linkage and Association Analysis with Exome Sequence Data Implicates *SLC25A40* in Hypertriglyceridemia

Elisabeth A. Rosenthal,[1,*] Jane Ranchalis,[1] David R. Crosslin,[2] Amber Burt,[1] John D. Brunzell,[3] Arno G. Motulsky,[1,2] Deborah A. Nickerson,[2] NHLBI GO Exome Sequencing Project, Ellen M. Wijsman,[1,2,4] and Gail P. Jarvik[1,2]

Hypertriglyceridemia (HTG) is a heritable risk factor for cardiovascular disease. Investigating the genetics of HTG may identify new drug targets. There are ~35 known single-nucleotide variants (SNVs) that explain only ~10% of variation in triglyceride (TG) level. Because of the genetic heterogeneity of HTG, a family study design is optimal for identification of rare genetic variants with large effect size because the same mutation can be observed in many relatives and cosegregation with TG can be tested. We considered HTG in a five-generation family of European American descent (n = 121), ascertained for familial combined hyperlipidemia. By using Bayesian Markov chain Monte Carlo joint oligogenic linkage and association analysis, we detected linkage to chromosomes 7 and 17. Whole-exome sequence data revealed shared, highly conserved, private missense SNVs in both *SLC25A40* on chr7 and *PLD2* on chr17. Jointly, these SNVs explained 49% of the genetic variance in TG; however, only the *SLC25A40* SNV was significantly associated with TG ($p = 0.0001$). This SNV, c.374A>G, causes a highly disruptive p.Tyr125Cys substitution just outside the second helical transmembrane region of the SLC25A40 inner mitochondrial membrane transport protein. Whole-gene testing in subjects from the Exome Sequencing Project confirmed the association between TG and *SLC25A40* rare, highly conserved, coding variants ($p = 0.03$). These results suggest a previously undescribed pathway for HTG and illustrate the power of large pedigrees in the search for rare, causal variants.

Introduction

Cardiovascular disease (CVD) is the leading cause of death in the United States and poses a significant morbidity and cost for treatment after cardiac events. CVD is associated with the correlated traits of high LDL, low HDL, high total cholesterol, high triglyceride (TG) (defined as $200 \leq TG < 500$ mg/dl in adults[1]), hypertension, diabetes, and metabolic syndrome. Furthermore, CVD is associated with environmental variables that can be confounded with lipid levels, such as obesity, poor diet, lack of exercise, and smoking.

Hypertriglyceridemia (HTG), defined as TG > 500 mg/dl in adults,[1] is a risk factor for CVD, independent of high LDL and low HDL.[2–7] Although HDL and TG levels are highly correlated, an independent role of HDL level in CVD etiology has been challenged by recent Mendelian randomization studies and the failure of cholesteryl ester transfer protein inhibitors to reduce vascular events.[8,9] Conversely, Mendelian randomization suggests a causal role of TG in CVD.[10] Elevated TG has been implicated in both microvascular and macrovascular endothelial damage with associated atherosclerosis.[6] Within the United States, ~16% of adults of European origin have high TG levels, indicating a need for further intervention.[7] However, studies of TG level and lipid metabolism have been difficult.[6,7] One reason for this difficulty is the existence of high within-individual variation of TG measurement that expands with increasing TG. High TG is also associated with high LDL and low HDL, making it difficult to tease apart the effect of specific lipids on CVD risk within studies.

There are currently few pharmacological treatments for elevated TG. The most common treatment, fibrates, effectively reduces elevated TG and reduces the risk for cardiovascular events.[11,12] Unfortunately, some 5% of individuals stop using fibrates because of side effects.[13] Other potential drugs, targeting different parts of the metabolic pathway, have been found to have intolerable complications such as fatty liver or to actually raise the risk of cardiovascular events.[13]

In order to find additional effective treatments, studies of TG need to be undertaken. Focusing on the genetic control of elevated TG may remove some of the confounding with LDL and HDL and lead to new drug targets. TG is known to be heritable and there are several known genetic mutations that influence TG levels, most notably those in the structural loci for ApoA5 and ApoC3.[14–21] In mice, expression of both *Apoa5* and *Apoc3* are associated with TG levels.[22–25] Whereas circulating levels of ApoA5 are negatively associated with TG levels, ApoC3 levels are positively correlated with TG. However, there is conflicting evidence in humans for an association between CVD and single-nucleotide variants (SNVs) within *APOC3* (MIM 606368) and *APOA5* (MIM 107720).[26–30] These and other known genetic variants explain only ~10% of the genetic

[1]Division of Medical Genetics, Department of Medicine, University of Washington, Seattle, WA 98195, USA; [2]Department of Genome Sciences, University of Washington, Seattle, WA 98195, USA; [3]Division of Metabolism, Endocrinology, and Nutrition, Department of Medicine, University of Washington, Seattle, WA 98195, USA; [4]Department of Biostatistics, University of Washington, Seattle, WA 98195, USA
*Correspondence: erosen@uw.edu
http://dx.doi.org/10.1016/j.ajhg.2013.10.019.

variation in TG,[20,21] which may explain the conflicting evidence indicating a relationship between regulatory SNVs and CVD.

The genetic heterogeneity in the etiology of high TG makes large family studies the optimal design for identification of novel TG loci with large effect sizes.[31] This design allows for the study of numerous people with an identical mutation and the ability to study cosegregation as well as association. For these reasons, we set out to discover genes underlying elevated TG in a single large family, described as familial combined hyperlipidemia (FCHL [MIM 144250]) in the 1980s and resampled in the 2000s, yielding five generations, four of which have phenotype and marker data. FCHL is genetically heterogeneous and characterized by variable atherogenic lipoprotein levels in multiple family members,[32–35] making it an ideal diagnosis to identify novel lipid level genes. This family includes four related individuals, including the proband, with TG measurement in the top 1/2 percentile of adults in the United States,[36] and previously reported genetic variation associated with TG does not explain the elevated levels. Segregation analyses supported at least one major gene for high TG in this family.

By using linkage analysis and whole-exome sequence data, we detected two candidate genes of interest. Although whole-exome sequencing has proven powerful in discovering underlying loci for simple Mendelian traits and de novo mutations, it has not yet been as successful in uncovering major contributing loci to complex traits. With family data, we have the opportunity to observe the same very rare variant in multiple family members and to evaluate cosegregation of a single variant with a phenotype. In this paper, we show that reduction of the search space by using linkage analysis in a large family, and then focusing on rare exome variants, can be used to successfully discover multiple major contributing loci for a heterogeneous, complex quantitative trait. This allows single gene burden testing in a second cohort, eliminating the need for multiple gene tests, which increases the power to detect an association.

Subjects and Methods

Subjects and Phenotype

The family presented here was ascertained in the 1980s as part of a cohort of four families of European American descent that had the diagnosis of FCHL. Observed TG in these families was previously shown to be associated with increased risk of CVD mortality.[37] The proband had fasting triglycerides >2,900 mg/dl[38] and normal LDL-C levels. The family was subsequently ascertained and diagnosed with FCHL, based on hypertriglyceridemia and hypercholesterolemia in first-degree relatives. The proband was originally hypothesized to represent a homozygote for FCHL, thought to be a Mendelian trait at that time. Re-collection of samples as well as sampling of additional family members, particularly in the youngest generation, took place in the 2000s. The final pedigree contains 5 generations of 121 individuals and includes five large sibships (n = 5, 5, 6, 7, and 9). Fasting TG was measured on 85 individuals (74 in the 1980s and 41 in the 2000s) via standard enzymatic methods, and reported in mg/dl.[39] The same methods were employed in the 1980s and 2000s. TG ranged between 22 and 2,926 mg/dl, with a median of 111 mg/dl for individuals not taking statin drugs. Twenty-six family members had TG > 200 mg/dl and 8 had TG > 500 mg/dl. Four related subjects in the second and third generations had TG > 1,000 mg/dl, making this family a good candidate for linkage analysis. In addition, none of the subjects were taking lipid-lowering medications in the 1980s and only nine were taking such medication in the 2000s. All subjects, or their representative, gave written, informed consent and this study was approved by the University of Washington Human Subject Review Board (FWA #00006878).

TG Adjustment

Fasting TG level is a highly right-skewed phenotype that is associated with age, sex, and body mass index in typical European American populations. Because of the ascertainment, the TG in this data set does not follow typical distributions and adjusting the TG via a linear model with the family data would not result in a reliable phenotype. To increase the power to detect linkage, we corrected for this ascertainment by adjusting TG (adjTG) by using the age- and sex-based means and standard deviations from The Lipid Research Clinics Population Studies Data Book, Volume 1,[40] added the constant 8 to avoid negative values, and then log transformed the data (log-adjTG) to reduce skew.

Because TG was measured in the 1980s and 2000s for 30 individuals, we generated three overlapping phenotype data sets. Data set "1980only" contains phenotypes from the 1980s only (n = 74, all statin-free). Data set "1980plus" contains phenotypes from the 1980s and phenotypes for only the newly sampled individuals from the 2000s who are not taking lipid-lowering medications (n = 85 statin-free). Data set "2000plus" contains all phenotypes from the 2000s for individuals not taking lipid-lowering medications and phenotypes from the 1980s for individuals who only have phenotype at this time point or who were taking statin drugs in the 2000s (n = 85 statin-free). Phenotypes in data sets 1980plus and 2000plus differ for 21 individuals.

Genotyping

The family was extensively genotyped via highly polymorphic markers, common SNVs, and next-generation exome sequencing. Quality control on all genotype data as well as alignment to the Rutgers Build 35 map was performed as described previously.[41,42] In brief, STR markers from Marshfield Panel 9 (1980s), Prevention Genetics Set 13 (2000s), and selected regions from Decode (1980s and 2000s) were genotyped in the family (n = 82). In addition, ~50K SNVs from the Illumina HumanCVD Bead chip were genotyped in the family (2000s, n = 64).[43] Finally, the exome was sequenced on the Solexa platform for 16 individuals selected for their phenotypes and informativeness for imputing the genotypes of others.[41,44] Novel SNVs of interest, detected with exome sequencing, were genotyped in the entire pedigree via custom TaqMan Genotyping Assays with the Applied Biosystems 7900HT real-time PCR system[45] or Veracode 384-plex Bead Plates.[46]

Joint Segregation and Linkage Analysis

Given that log-adjTG is a complex trait with an unknown mode of inheritance in these data, we used an iterative Bayesian oligogenic

Table 1. Percent Variance of log-adjTG Explained in the Family, by Previously Reported SNVs

Chr	Gene (SNV, Proxy)	MIM #	SNV	Proxy	% Vtot (SD)	% Vg (SD)	p Value
1	*DOCK7*	604774	rs2131925	rs1748195	3 (3)	14 (16)	0.84
1	*GALNT2*	602274	rs1321257	rs4846914	4 (3)	23 (20)	0.15
2	*APOB*	107730	rs1042034	NA	4 (3)	16 (17)	0.73
2	*GCKR*	600842	rs1260326	NA	4 (3)	18 (19)	0.09
3	*MSL2,*[a] *PCCB*	614802, 232050	rs645040	rs3821445	3 (2)	17 (19)	0.34
4	*AFF1*	159557	rs442177	rs3775214	2 (2)	13 (17)	0.68
7	*TBL2*	605842	rs17145738	NA	5 (4)	20 (18)	0.04
8	*LPL,*[a] *LPL*	609708	rs12678919	rs12679834	2 (2)	14 (19)	0.64
8	*TRIB1*[a]	609461	rs2954029	rs17321515	3 (2)	13 (16)	0.60
10	*CYP26A1*[a]	602239	rs2068888	rs4418728	4 (3)	24 (21)	0.25
11	*FADS1, FADS2*	606148, 606149	rs174546	rs1535	3 (2)	14 (18)	0.82
11	*APOA5*	606368	rs3135506	NA	2 (2)	14 (18)	0.67
15	*LIPC*	151670	rs6074	rs871804	4 (3)	18 (17)	0.17
15	*LIPC*	151670	rs261342	NA	2 (2)	15 (19)	0.94
16	*CTF1*[a]	600435	rs11649653	NA	3 (3)	14 (16)	0.47
16	*CETP*	118470	rs7205804	rs1532625	3 (3)	21 (21)	0.99
19	*APOE,*[a] *APOC1*[a]	107741, 107710	rs439401	NA	5 (3)	20 (18)	0.12
19	*SUGP1*, *GATAD2A*	607992, 614997	rs10401969	rs3794991	1 (2)	10 (17)	0.70
20	*PLTP*	172425	rs4810479	NA	7 (5)	30 (18)	0.56

Estimated percent total variance (Vtot), percent genetic variance (Vg), and p value for pedigree adjusted association of log-adjTG and known or suspected pathogenic SNVs in the family. Proxy SNVs from the CVD chip were used in place of the reported SNVs that did not exist on the CVD chip. Proxy SNVs are correlated with the original SNV ($r^2 > 0.8$). *APOA5* (rs3135506) was genotyped separately in the family. The second column from left gives the gene name for the SNV and its proxy, if necessary. Gene names are from NCBI and aliases are as follows: *DOCK7* = *ANGPTL3*, *AFF1* = *KHL8*, *TBL2* = *MLXIPL*, *SUGP1* = *CILP2*. SD indicates standard deviation.
[a]SNV is near but not within a gene.

joint segregation and linkage analysis approach from the package Loki2.4.7.[47] This approach has many advantages for these data. First, Loki is capable of handling multipoint analysis on a large pedigree. Second, the user does not supply a trait model; rather, a prior distribution on the genetic effects is supplied, allowing the trait model, including the number of QTLs, to vary within a plausible model space. Third, missing trait-locus genotypes can be sampled at each iteration, given the marker segregation in the pedigree. Fourth, genotype covariates can be included in the model, allowing for estimation of the genetic variation explained by a SNV at each iteration, including the information from the estimated missing genotypes. Over an entire run, the average total variation (Vtot) and genetic variation (Vg) explained can then be used to estimate the percent Vtot and Vg explained by that SNV. In addition to modeling the possibility of multiple QTLs, variance resulting from any polygenic effects is captured by Vg. Because calculating p values is time consuming with this method, we calculated the Bayes factor (BF) for 2 cM intervals across the genome, which compares the posterior odds to the prior odds that a QTL is located in the region. We consider a maxBF > 25 as suggestive of linkage.[41,48] All identified linkage regions were jointly analyzed, allowing the regions to compete for the modeled QTLs.

Because the trait under study is skewed, it is important to specify a prior distribution that is expansive enough to include a likely model but not so broad that the space cannot be adequately sampled. Furthermore, the detected regions of linkage should be consistent given differing prior distributions. By using methods described previously,[49–51] we used several values for the underlying additive genetic effects, ranging between 1 and 8 times the standard deviation (SD) of log-adjTG in order to verify consistency of detected linkage regions. Here we report the results for the prior distribution with mean additive genetic effect = 2.7 times the SD of log-adjTG. All runs contained 1,000 burn-in iterations followed by 200,000 iterations (1 chromosome) or 500,000 iterations (2 chromosomes jointly), of which every 10^{th} iteration was saved. The prior distribution on the number of QTLs followed a curtailed Poisson(1) distribution with a maximum of 15. The total map length was assumed to be 3,000 cM. The prior distribution on the allele frequency for all QTLs is set to Uniform(0,1) and cannot be changed by the user.

Candidate Locus Analysis

Candidate loci were assessed for the estimated percent Vtot and Vg that they explained in the family. These candidate variants were chosen from two sources (Tables 1 and 2 and Table S1 available online). First, candidate SNVs known or thought to be associated with TG,[21,52] including *APOA5* variant rs3135506, were assessed within the family primarily via the CVD chip data. If the reported SNV was not on the CVD chip, we attempted to find a correlated

Table 2. Distribution of Novel SNVs under the Linkage Signals on Chr7 and Chr17

Chr.	7	17
# novel sites	53	20
Intergenic	2	1
Intronic	4	1
3′ UTR	1	1
5′ UTR	1	0
Synonymous	23	2
Splice	0	1
Missense	22	14
GERP > 3	12	6
Shared	1	4
Liver expressed	1	2

The final row indicates the number of novel SNVs that are missense or splice sites, have Genomic Evolutionary Rate Profiling score (GERP) > 3, and are shared by at least two relatives with high log-adjTG. Novel is defined as not existing in dbSNP134.

proxy SNV ($r^2 > 0.8$) within 500 bases of the reported SNV.[53,54] *APOA5* SNV rs3135506 was genotyped separately. A Bonferroni adjustment was used for the testing of these known candidate loci.

Candidate rare SNVs were chosen from the exome data in the linkage regions. Exome sites were filtered based on novelty, coding effect, conservation across mammals, and sharing of the rare allele among the four individuals with TG > 1,000. Because the genetic basis of TG in this family had yet to be discovered, we hypothesized that the high levels of TG were due to novel variation. We defined a variant as novel if it did not have an rsID in dbSNP134 at the time of exome sequencing (April 2012)[55] and occurred at most once in the NHLBI Exome Sequencing Project (ESP) data on 6,500 exomes. Only sites that had an effect on the coding sequence (nonsense, missense, or splice) were kept. Conservation across mammals was calculated by the Genomic Evolutionary Rate Profiling (GERP) score;[56] positive values indicate conservation and negative values indicate lack of conservation. Only sites with a GERP > 3 were considered because we expect only mutations at locations with high conservation across mammals would be pathogenic for high TG. Finally, the rare allele had to exist in at least two of the four individuals with TG > 1,000. All pedigree members were genotyped for the novel sites selected with these criteria to confirm their estimated effect. Variance components analysis in the package SOLAR[57] was used to calculate p values for association, in a mixed model analysis that adjusted for correlation among related individuals through the kinship matrix, which captures any underlying shared polygenic effect.

Validation of Novel Gene Association

We sought confirmation of association with TG for any gene containing a novel SNV that remained associated with the trait after genotyping of the full pedigree. By using TG data from unrelated individuals in the PennCATH, TRIUMPH, Cleveland Clinic, ARIC, CARDIA, CHS, FHS, JHS, MESA, and WHI cohorts in the dbGaP posted ESP data, we combined all rare (MAF < 0.5%) missense, nonsense, and splice SNVs with GERP > 4.8 into a single genotype factor (1 = presence of a minor allele and 0 = absence of all minor alleles) in individuals with measured TG. We used a GERP cutoff of 4.8 because this is the 75th percentile for GERP of all identifiable LDL-raising pathogenic SNVs in *LDLR* (MIM 606945)[55,56,58,59] (data not shown). We assumed that individuals missing genotypes at these rare SNVs were noncarriers. We performed this two-sided whole-gene test via a linear model for log(TG) on the {0, 1} genotype factor, using the package R with and without adjusting for race (first three principal components or ethnic group), age, and sex. Statin and fibrate usage is not contained in this data. However, because these medications reduce TG levels, their absence from the model serves to make the test more conservative.

Results

Phenotype

The adjTG is a stable phenotype over time in this data set. The correlation of adjTG values ~20 years apart, in the 21 individuals with data at both time points and not on statins, is highly significant ($r^2 = 0.78$, $p = 2.7 \times 10^{-5}$) with positive slope coefficient ($\beta = 0.5$). adjTG level appears to have larger variance at higher levels, as expected. The adjusted values were independent of age and sex, as expected.

Known TG SNVs

There were 18 SNVs, or their proxies, on the CVD chip of the total 35 (Table 1) that were reported to have an effect on TG.[21,52] These 18 SNVs and the *APOA5* SNV, rs3135506, explained at most 7% Vtot in log-adjTG in this family (Table 1). In addition, there was no association with the cumulative genetic risk score[21] and log-adjTG ($p = 0.20$). rs4810479, in the 5′ region of *PLTP* (MIM 172425), explained the most Vtot in log-adjTG, 7% (SD = 5%), but was not significantly associated with log-adjTG ($p = 0.56$), adjusting for pedigree structure. The *APOA5* SNV, rs3135506, explained 2.1% Vtot (SD = 2%). SNV rs17145738 in *TBL2* ([MIM 605842] alias *MLXIPL*) and near one of our linkage signals (see below), explained 5% Vtot in log-adjTG (SD = 4%) and had the lowest p value for association, $p = 0.04$, adjusting for pedigree structure. This result is not significant given the Bonferroni adjustment for multiple testing (Bonferonni cutoff of 0.003). Furthermore, the minor allele is associated with lower log-adjTG, indicating that this common SNV is unlikely to explain the high TG observed in this family. Because these known SNVs did not explain much of the Vtot in log-adjTG and were not significantly associated with log-adjTG, we performed linkage analysis to identify regions of interest.

Joint Segregation and Linkage Analysis Results

Segregation analysis, both alone and in conjunction with linkage analysis, shows evidence of at least one QTL underlying log-adjTG. Given the prior mean additive genetic effect = 2.7 times the SD of log-adjTG, the posterior probability of at least one QTL is 0.97, 0.88, and 0.998 for data

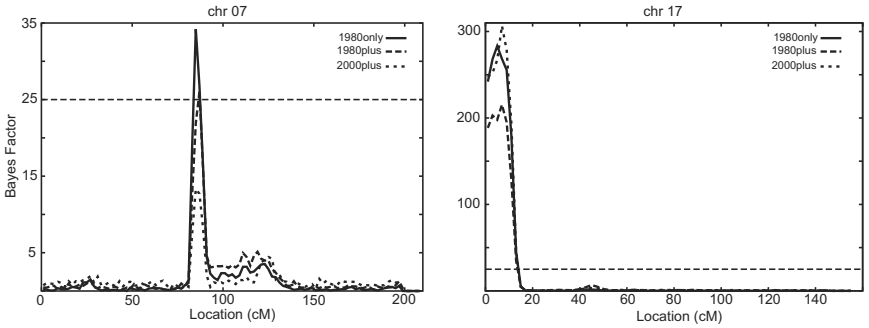


Figure 1. Evidence of Linkage, as Measured by the Bayes Factor, on Chromosomes 7 and 17 for the Three Data Sets 1980only, 1980plus, and 2000plus
Location is given in cM and spans the entire length of the chromosomes. The horizontal dashed line at 25 indicates the cutoff for evidence for linkage.

sets 1980only, 1980plus, and 2000plus. The posterior probability of exactly one QTL is 0.83, 0.54, and 0.23 for data sets 1980only, 1980plus, and 2000plus. In addition, the three overlapping data sets show support for log-adjTG linkage to chr7 and chr17 (Figure 1). Data sets 1980only and 1980plus support linkage to both chromosomes. Data set 2000plus shows support for only chr17. Using differing prior distributions on the genetic effects did not change the locations of the linkage signals (data not shown). Because data set 1980plus shows support for both regions and contains more phenotype data than does 1980only, we continued the analysis with this data set. The posterior predicted model for 1980plus is a recessive mode of inheritance.

Novel SNVs

Within and near these linkage regions, we found five novel SNVs that met our criteria for candidacy, labeled N7, N17a, N17b, N17c, and N17d, that were private to this family (Tables 2 and 3). Variant N7 is shared by all four IDs with TG > 1,000. Although chr17 contained four SNVs of interest, they were shared by three different pairs of IDs with TG > 1,000: N17a at one pair, N17b at another, and N17c and N17d at a different pair. These five SNVs explain 19%, 7.8%, 4.8%, 5.3%, and 5.3% of Vtot, respectively, in data set 1980plus. It is expected that the estimated %Vtot explained will decrease upon full genotyping in the family, because the most informative people have measured or imputable genotypes and the branch of the family without variability of the SNV of interest will lower the Vtot explained. Therefore, we chose to pursue full genotyping of N7 and N17a only because they explain >7% of Vtot with just the exome genotypes, which is more than the known SNVs in *PLTP*, *APOA5*, and *TBL2* noted above. These SNVs were in *SLC25A40* (MIM 610821) and *PLD2* (MIM 602384), respectively. Additionally, no known pathogenic or novel likely pathogenic variation was found in the known TG genes *LPL* (MIM 09708), *APOC2* (MIM 608083), *APOA5*, *GPIHBP1* (MIM 612757), and *LMF1* (MIM 611761) or in the lipid-related genes *CETP* (MIM 118470) and *LDLR* in the 16 exomes.

Jointly analyzing the chromosomes and their associated novel SNVs, genotyped in the full family, clarifies the relationship among these two regions (Figures 2 and S1). As expected, given the complex model, adjustment for each SNV impacts the linkage evidence for the other region.

Table 3. Percent Variance of log-adjTG Explained in the Family, by Candidate Novel SNVs

Variant	N7	N17a	N17b	N17c	N17d
Chr	7	17	17	17	17
Pos.	87477251	4711152	6021379	2290581	4496467
Gene	*SLC25A40*	*PLD2*	*WSCD1*	*MNT*	*SMTNL2*
RefSeq	NM_018843.3	NM_002663.4	NM_015253.1	NM_020310.2	NM_001114974.1; NM_198501.2
NT change	c.374A>G	c.85A>T	c.1246A>G	c.1363G>C	c.730+1G>T; c.298+1G>T
Protein change	p.Tyr125Cys	p.Thr29Ser	p.Arg416Gly	p.Val455Leu	NA
Coding effect	missense	missense	missense	missense	splice
GERP	5.1	5.1	4.3	3.9	5.0
%Vtot as exome variant	19	7.8	4.8	5.3	5.3
%Vg as exome variant	49	23	22	27	27
maxBF	1.8	48	103	72	66

Estimated percent total variance (Vtot) and genetic variance (Vg) in log-adjTG explained by each of the candidate novel SNVs within the linkage regions, using only individuals that have exome data. MaxBF are given for when the novel variant is included as a marker and covariate in the joint linkage and association analysis. Missing genotype data are imputed for other family members. Physical positions are from build 37. In dbSNP137, rsIDs have been assigned to SNVs N17b, N17c, and N17d as follows: N17b = rs200724890, N17c = rs201365025, N17d = rs202160684. There are two RefSeq numbers for *SMTNL2* because it has two isoforms.

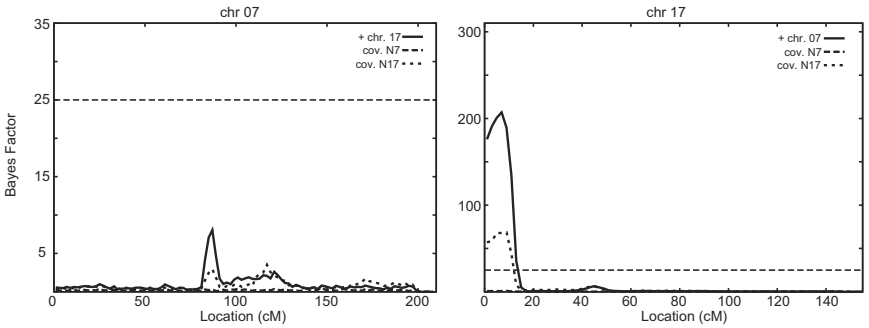


Figure 2. Evidence of Linkage in Data Set 1980plus, as Measured by the Bayes Factor, on Chromosomes 7 and 17 when Analyzed Jointly, and with either SNV N7 or N17 as a Genotype Covariate
Location is given in cM and spans the entire length of the chromosomes. The horizontal dashed line at 25 indicates the original cutoff for evidence for linkage.

When both chromosomes are analyzed jointly, the linkage evidence on chr7 was reduced (maxBF = 8), but the linkage to chr17 remained (maxBF = 207). Linkage results are similar when including the fully genotyped SNVs N7 and N17a as markers only, in the analysis, indicating that when neither of the SNVs is included as a genotype covariate, the posterior distribution favors chr17 over chr7. However, only SNV N7 explained all evidence for linkage (chr17 maxBF reduced from 207 to 1, and chr7 maxBF reduced from 8 to 0.5) when included as a genotype covariate (Figure 2). When included as a genotype covariate, N17a explained some of the signal on chr17 (maxBF decreases from 207 to 69) and chr7 (maxBF reduces from 8 to 3.5).

The relative strength of each candidate SNV is reflected in the percent Vg explained by each, as well as the significance of the pedigree-adjusted association with log-adjTG. N7 (Ngt = 40) explained 19% of log-adjTG Vtot and 49% of Vg. N17a (Ngt = 78) explained 7% of log-adjTG Vtot and 25% Vg. When both variants are included jointly, they explained 49% of log-adjTG Vg; with N7 and N17a explaining 40% (p = 0.0001) and 8.6% (p = 0.31) of log-adjTG Vg, respectively. We also explored the significance of the known TG SNVs while adjusting for N7 genotype. In this case, each SNV explained <5% of Vtot and were not statistically significant (unadjusted p > 0.08). The *TBL2* SNV remained insignificant (unadjusted p = 0.08). Furthermore, this SNV is carried by individuals with log-adjTG below the mean, and these are not the same individuals as those that carry the N7 variant. Although N7 is associated with log-adjTG, it is not associated with LDL-C in this family (p = 0.83, Figure S2).

Table 4. Demographic Count Data and Age Distribution for the ESP Cohort, Broken down by Quantile of log(TG)

Log(TG) quantile	M	F	EA	AA	Mean (Min, Max) Age
[1.77,4.36]	341	597	434	504	50.15 (12, 85)
(4.36,4.74]	298	666	461	503	57.60 (18, 84)
(4.74,5.13]	312	596	522	386	58.27 (20, 93)
(5.13,6.84]	343	589	664	268	59.69 (21, 90)

This data set contains 3,770 verified unrelated individuals. Abbreviations are as follows: M, male; F, female; EA, European American; AA, African American.

Confirmation of *SLC25A40* Association

SLC25A40 was significantly associated with log(TG) in a separate cohort by using ESP data (Table 4). There were five rare missense SNVs (Table 5) in *SLC25A40* with GERP > 4.8 identified in ESP subjects with TG data. Notably, the *SLC25A40* SNV in the linkage family had a GERP of 5.1 and each of these five SNVs had GERPs > 5.1; two were predicted to be "probably damaging" by PolyPhen.[60] Six subjects of European American (EA) descent and two of African American (AA) descent carried one of these variants. The two of AA descent carried the same SNV; this SNV did not vary in EA in the entire ESP data. The SNVs found in the EA-descent subjects were also specific to that ancestry group. No subject carried more than one of these *SLC25A40* SNVs. After pooling carrier status for each of these SNVs into a single factor, the whole-gene test gives a significant positive association between these high GERP variants at *SLC25A40* and log(TG) (β = 0.45, two-sided p = 0.02), adjusted for age and sex. This whole-gene association with log(TG) remains significant and positive when adjusted for race, age, and sex (β = 0.42, two-sided p = 0.03). Results are similar when applied to individuals older than 25 and when removing individuals with a rare SNV in the gene, whose GERP score ranges between 2 and 4.8. Although the sample size is too small for stratified testing, the coefficient is positive in both ancestry groups (β = 0.18 and 1.1 for EA and AA descent, respectively). Although we focused on *SLC25A40* as the best candidate from this region, we sought to rule out associations of other regional genes and log(TG) in the ESP cohort by similar analyses (Table S2). Carrier status at *PLD2* (p = 0.48), *WSCD1* (p = 0.19), *MNT* (p = 0.81 [MIM 603039]), and *SMTNL2* (p = 0.87) did not predict log(TG).

Discussion

We identified two genes, *SLC25A40* and *PLD2*, each containing a novel variant that cosegregates with severely high TG levels in a single large family ascertained for segregation of FCHL. Joint analysis of the linkage regions and these variants indicated that the variant in *SLC25A40* is the most probable variant for pathogenicity. However, we

Table 5. Rare SNVs with GERP > 4.8 in *SLC25A40* Found in Individuals in the ESP Data

Position	rsID	NT Change	MAF % (EA/AA/All)	AA Change	GERP	PolyPhen	n
87470986	NA	c.785A>G	0.0118/0.0/0.0078	p.Gln262Arg	5.92	B	1 EA
87483601	NA	c.182G>T	0.0/0.0454/0.0154	p.Gly61Val	5.91	PD	2 AA
87473169	rs145515966	c.641G>C	0.0117/0.0/0.0077	p.Trp214Ser	5.67	PD	1 EA
87473143	rs140104130	c.667T>A	0.0349/0.0/0.0231	p.Trp223Arg	5.67	B	3 EA
87477276	NA	c.349G>A	0.0116/0.0/0.0077	p.Ala117Thr	5.13	B	1 EA

MAF % based on 8,600 European (EA) and 4,406 African American (AA), self-reported ancestry subjects; TG were measured in 2,168 EA and 1,793 AA subjects, with ancestry genetically verified. Abbreviations are as follows: B, benign; PD, probably damaging.

cannot rule out a possible interaction or additive effect between these and other identified variants. No known common pathogenic variants were found to be associated with TG in this family. We were able to replicate the effects of rare, conserved missense mutations only in *SLC25A40* in the separate ESP cohort.

SLC25A40 is solute carrier family 25, member 40, and its product localizes to the mitochondria, where it is involved in membrane transport.[61] Although the gene is ubiquitously expressed, its product is found mainly in the adrenal gland.[62] The novel variant in this gene (RefSeq accession number NM_018843.3), c.374A>G, is located in the seventh exon (of 12), is highly conserved (GERP = 5.1), and causes a p.Tyr125Cys substitution in the 338 amino acid protein. This change is just outside the second helical transmembrane region of the protein. Although there is no PolyPhen prediction, the tyrosine at amino acid 125 is perfectly conserved across vertebrates.[58,63] Furthermore, the Grantham score is 194, indicating a high chemical dissimilarity from the wild-type.[64] This particular mutation causes a new cysteine that neighbors a cysteine at position 124 and increases the number of cysteines to 9. Cysteines play a critical role in protein folding because they are the only amino acids that can create disulfide bonds that stabilize the folded form of the protein.[65] This change of a single amino acid to a cysteine allows for an increased number of folding possibilities, and possibly a lower entropy stabilization, which can disrupt the protein's function.

It is possible that this specific variant is not pathogenic, but is in linkage disequilibrium with a causal variant not detected by our exome sequencing. Further gene sequencing, as well as functional studies, are needed to verify the association. However, we found evidence for a pathogenic effect of high TG for rare, conserved (GERP > 5) missense variants in *SLC25A40* in a separate, unrelated cohort, supporting the idea that *SLC25A40* affects TG levels.

Although our analyses suggest that a *PLD2* missense variant may also impact TG, whole-gene testing in a separate cohort does not support pathogenicity for the gene. Further study may be warranted, because this gene may have biological relevance to TG. *PLD2* is phospholipase D2, whose product catalyzes the hydrolysis of phosphatidylcholine to phosphatidic acid and choline, and is ubiquitously expressed. The novel variant in this gene (RefSeq NM_002663.4), c.85A>T, causes a p.Thr29Ser (out of 934 amino acids) and is highly conserved across mammals.

Three other candidate genes identified on chr17, *WSCD1*, *MNT*, and *SMTNL2*, also were not significantly associated with TG in this separate cohort. Of these three genes, only *SMNTL2* is expressed in liver. *SMNTL2* is smoothelin-like 2, of which very little is currently known. Because neither *WSCD1* nor *MNT* are expressed in liver, they are less likely to be involved in TG regulation. Little is known about *WSCD1*, the WSC domain containing 1 gene. *MNT* encodes the MAX dimerization protein and is thought to repress transcription by binding to DNA binding proteins.[66,67] Reliance on the exome-sequence data rather than full-genome sequencing may have missed other potentially causal variants.

It is worth noting that the size of the linkage signals do not necessarily correlate with the candidate gene with the most evidence for causality. When chr7 and chr17 are analyzed jointly, the posterior distribution prefers a recessive QTL with moderate allele frequency on chr17. This is most probably due to the ascertainment scheme, lack of prior information on the allele frequency, and the observed phenotype segregation. The ascertainment scheme, which is useful for observing a rare allele multiple times, artificially inflates the observed allele frequency. In addition, the uniform prior distribution on the allele frequencies does not allow for a preference for rare alleles. Finally, the skew of the phenotype data in which offspring may have much higher phenotype than their parents results in evidence for a recessive trait. Therefore, segregation analysis showed overwhelming support for a recessive gene with a common allele frequency, which fit the marker and trait segregation at chr17. Further analysis including the exome-sequence data as covariate effects, which included a fixed rare allele frequency in the model, favored a dominant gene as evidenced by the high association of the SNV on chr7 with TG in the family.

The sampling and analysis design used here is a powerful approach for successful identification of novel genes and biological pathways underlying heterogeneous complex traits. Ascertainment of a large family segregating extreme values of a quantitative phenotype reduces the number of

putative underlying highly penetrant loci, increasing the chances of finding a variant with biological relevance in multiple family members. The large size of the pedigree, although a fraction of the size of a typical GWAS sample, provides multiple dimensions of information that, when used jointly, substantially increases the power to detect linkage.[31] Furthermore, correction for the ascertainment of these extreme phenotypes, as in this study, further increases power to detect linkage. The use of linkage analysis further drastically reduces the search space, limiting the multiple testing problem seen with GWASs or exome-wide burden tests. Finally, joint linkage and association in such pedigrees, as when both chromosomes 7 and 17 compete for the modeled QTL and the candidate novel variants are included as covariates, can help determine which of the identified candidate genes has the most promise. One limitation of this approach is that the same rare variant may not be found in a second family or enough unrelated individuals to provide confirmation; however, a single or limited number of gene-based burden tests can be made in an unrelated sample, again avoiding the penalty of exome-wide multiple comparisons.

Although common alleles have been found to explain a small portion of moderate to severe HTG in unrelated samples, we show that rare variation plays a major role in explaining severe HTG in a family previously diagnosed with FCHL and that this knowledge can be used to identify novel genes and biological pathways of interest. Additionally, our linkage results and the replication of a HTG effect in the ESP data, which is not ascertained on FCHL, lend support to an oligogenic inheritance of the lipid traits contributing to the diagnosis of FCHL in this family. Further evidence for this includes the fact that LDL-C level is not associated with the *SLC25A40* variant in these data, the proband had normal LDL-C and apoB levels, we do not detect linkage between LDL-C and this region of chr7 (data not shown) by similar methodology, and we have previously reported linkage between apoB level and chr4 in this family.[42] FCHL families may not have a monogenic disorder as earlier described, but rather a confluence of separate traits in the same family. Furthermore, the complex nature of the trait within this family is borne out by the fact that three carriers of the SLC25A40 variant do not have HTG, although two of them have TG near the 95th percentile (Figure S2), and that one individual with TG $>$ 500 does not carry the SLC25A40 variant, indicating that other genetic variants may be influential. However, the polygenic effects are accounted for in the analysis and have a small impact on the trait, relative to the major gene component. Indeed, *SLC25A40* may be shown to cause familial hypertriglyceridemia (MIM 145750), given that it does not appear to raise LDL-C. We note that our proband had highly elevated TG for a subject with FCHL, which is why the proband was thought to potentially be a FCHL homozygote; thus, the relationship of this locus to a more typical FCHL family with more modest HTG is not yet clear. Neither *SLC25A40* nor *PLD2* (or their respective biological pathways) have been previously implicated in triglyceride levels, to our knowledge. Functional studies, in vitro and in vivo, will need to be carried out to verify that they impact TG levels and to discover the mechanism by which they have an effect. Given the relevance of high TG to CVD, TG-lowering treatments targeted to these pathways may be identified.

Supplemental Data

Supplemental Data include Supplemental Acknowledgments, two figures, and two tables and can be found with this article online at http://www.cell.com/AJHG/.

Acknowledgments

Thanks go to Peter Byers for his helpful comments. Funding for this analysis was provided by National Institutes of Health grants P01 HL030086, T32 GM007454, and R01 HL094976 and the State of Washington Life Sciences Discovery Fund award to the Northwest Institute of Genetic Medicine (grant 265508). The authors wish to acknowledge the support of the National Heart, Lung, and Blood Institute (NHLBI) and the contributions of the research institutions, study investigators, field staff, and study participants in creating this resource for biomedical research. Funding for GO ESP was provided by NHLBI grants RC2 HL-103010 (HeartGO), RC2 HL-102923 (LungGO), and RC2 HL-102924 (WHISP). The exome sequencing was performed through NHLBI grants RC2 HL-102925 (BroadGO) and RC2 HL-102926 (SeattleGO). Genotyping services were provided through the RS&G Service by the Northwest Genomics Center at the University of Washington, Department of Genome Sciences, under US Federal Government contract number HHSN268201100037C from the National Heart, Lung, and Blood Institute.

Received: July 25, 2013
Revised: September 12, 2013
Accepted: October 21, 2013
Published: November 21, 2013

Web Resources

The URLs for data presented herein are as follows:

NHLBI Exome Sequencing Project (ESP) Exome Variant Server, http://evs.gs.washington.edu/EVS/
Online Mendelian Inheritance in Man (OMIM), http://www.omim.org/
R statistical software, http://www.r-project.org/
RefSeq, http://www.ncbi.nlm.nih.gov/RefSeq

References

1. National Cholesterol Education Program (NCEP) Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults (Adult Treatment Panel III) (2002). Third Report of the National Cholesterol Education Program (NCEP) Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults (Adult Treatment Panel III) final report. Circulation *106*, 3143–3421.

2. Hokanson, J.E., and Austin, M.A. (1996). Plasma triglyceride level is a risk factor for cardiovascular disease independent of high-density lipoprotein cholesterol level: a meta-analysis of population-based prospective studies. J. Cardiovasc. Risk *3*, 213–219.
3. Patel, A., Barzi, F., Jamrozik, K., Lam, T.H., Ueshima, H., Whitlock, G., and Woodward, M.; Asia Pacific Cohort Studies Collaboration (2004). Serum triglycerides as a risk factor for cardiovascular diseases in the Asia-Pacific region. Circulation *110*, 2678–2686.
4. Tirosh, A., Rudich, A., Shochat, T., Tekes-Manova, D., Israeli, E., Henkin, Y., Kochba, I., and Shai, I. (2007). Changes in triglyceride levels and risk for coronary heart disease in young men. Ann. Intern. Med. *147*, 377–385.
5. Miller, M., Cannon, C.P., Murphy, S.A., Qin, J., Ray, K.K., and Braunwald, E.; PROVE IT-TIMI 22 Investigators (2008). Impact of triglyceride levels beyond low-density lipoprotein cholesterol after acute coronary syndrome in the PROVE IT-TIMI 22 trial. J. Am. Coll. Cardiol. *51*, 724–730.
6. Kohli, P., and Cannon, C.P. (2012). Triglycerides: how much credit do they deserve? Med. Clin. North Am. *96*, 39–55.
7. Miller, M., Stone, N.J., Ballantyne, C., Bittner, V., Criqui, M.H., Ginsberg, H.N., Goldberg, A.C., Howard, W.J., Jacobson, M.S., Kris-Etherton, P.M., et al.; American Heart Association Clinical Lipidology, Thrombosis, and Prevention Committee of the Council on Nutrition, Physical Activity, and Metabolism; Council on Arteriosclerosis, Thrombosis and Vascular Biology; Council on Cardiovascular Nursing; Council on the Kidney in Cardiovascular Disease (2011). Triglycerides and cardiovascular disease: a scientific statement from the American Heart Association. Circulation *123*, 2292–2333.
8. Voight, B.F., Peloso, G.M., Orho-Melander, M., Frikke-Schmidt, R., Barbalic, M., Jensen, M.K., Hindy, G., Hólm, H., Ding, E.L., Johnson, T., et al. (2012). Plasma HDL cholesterol and risk of myocardial infarction: a mendelian randomisation study. Lancet *380*, 572–580.
9. Schwartz, G.G., Olsson, A.G., Abt, M., Ballantyne, C.M., Barter, P.J., Brumm, J., Chaitman, B.R., Holme, I.M., Kallend, D., Leiter, L.A., et al.; dal-OUTCOMES Investigators (2012). Effects of dalcetrapib in patients with a recent acute coronary syndrome. N. Engl. J. Med. *367*, 2089–2099.
10. Sarwar, N., Sandhu, M.S., Ricketts, S.L., Butterworth, A.S., Di Angelantonio, E., Boekholdt, S.M., Ouwehand, W., Watkins, H., Samani, N.J., Saleheen, D., et al.; Triglyceride Coronary Disease Genetics Consortium and Emerging Risk Factors Collaboration (2010). Triglyceride-mediated pathways and coronary disease: collaborative analysis of 101 studies. Lancet *375*, 1634–1639.
11. Lee, M., Saver, J.L., Towfighi, A., Chow, J., and Ovbiagele, B. (2011). Efficacy of fibrates for cardiovascular risk reduction in persons with atherogenic dyslipidemia: a meta-analysis. Atherosclerosis *217*, 492–498.
12. Bruckert, E., Labreuche, J., Deplanque, D., Touboul, P.J., and Amarenco, P. (2011). Fibrates effect on cardiovascular risk is greater in patients with high triglyceride levels or atherogenic dyslipidemia profile: a systematic review and meta-analysis. J. Cardiovasc. Pharmacol. *57*, 267–272.
13. Wierzbicki, A.S., Hardman, T.C., and Viljoen, A. (2012). New lipid-lowering drugs: an update. Int. J. Clin. Pract. *66*, 270–280.
14. Pennacchio, L.A., Olivier, M., Hubacek, J.A., Krauss, R.M., Rubin, E.M., and Cohen, J.C. (2002). Two independent apolipoprotein A5 haplotypes influence human plasma triglyceride levels. Hum. Mol. Genet. *11*, 3031–3038.
15. Talmud, P.J., Hawe, E., Martin, S., Olivier, M., Miller, G.J., Rubin, E.M., Pennacchio, L.A., and Humphries, S.E. (2002). Relative contribution of variation within the APOC3/A4/A5 gene cluster in determining plasma triglycerides. Hum. Mol. Genet. *11*, 3039–3046.
16. Romeo, S., Pennacchio, L.A., Fu, Y., Boerwinkle, E., Tybjaerg-Hansen, A., Hobbs, H.H., and Cohen, J.C. (2007). Population-based resequencing of ANGPTL4 uncovers variations that reduce triglycerides and increase HDL. Nat. Genet. *39*, 513–516.
17. Kathiresan, S., Melander, O., Guiducci, C., Surti, A., Burtt, N.P., Rieder, M.J., Cooper, G.M., Roos, C., Voight, B.F., Havulinna, A.S., et al. (2008). Six new loci associated with blood low-density lipoprotein cholesterol, high-density lipoprotein cholesterol or triglycerides in humans. Nat. Genet. *40*, 189–197.
18. Kooner, J.S., Chambers, J.C., Aguilar-Salinas, C.A., Hinds, D.A., Hyde, C.L., Warnes, G.R., Gómez Pérez, F.J., Frazer, K.A., Elliott, P., Scott, J., et al. (2008). Genome-wide scan identifies variation in MLXIPL associated with plasma triglycerides. Nat. Genet. *40*, 149–151.
19. Teslovich, T.M., Musunuru, K., Smith, A.V., Edmondson, A.C., Stylianou, I.M., Koseki, M., Pirruccello, J.P., Ripatti, S., Chasman, D.I., Willer, C.J., et al. (2010). Biological, clinical and population relevance of 95 loci for blood lipids. Nature *466*, 707–713.
20. Nilsson, S.K., Heeren, J., Olivecrona, G., and Merkel, M. (2011). Apolipoprotein A-V; a potent triglyceride reducer. Atherosclerosis *219*, 15–21.
21. Johansen, C.T., Kathiresan, S., and Hegele, R.A. (2011). Genetic determinants of plasma triglycerides. J. Lipid Res. *52*, 189–206.
22. Baroukh, N., Bauge, E., Akiyama, J., Chang, J., Afzal, V., Fruchart, J.C., Rubin, E.M., Fruchart-Najib, J., and Pennacchio, L.A. (2004). Analysis of apolipoprotein A5, c3, and plasma triglyceride concentrations in genetically engineered mice. Arterioscler. Thromb. Vasc. Biol. *24*, 1297–1302.
23. Ito, Y., Azrolan, N., O'Connell, A., Walsh, A., and Breslow, J.L. (1990). Hypertriglyceridemia as a result of human apo CIII gene expression in transgenic mice. Science *249*, 790–793.
24. Pennacchio, L.A., Olivier, M., Hubacek, J.A., Cohen, J.C., Cox, D.R., Fruchart, J.C., Krauss, R.M., and Rubin, E.M. (2001). An apolipoprotein influencing triglycerides in humans and mice revealed by comparative sequencing. Science *294*, 169–173.
25. Maeda, N., Li, H., Lee, D., Oliver, P., Quarfordt, S.H., and Osada, J. (1994). Targeted disruption of the apolipoprotein C-III gene in mice results in hypotriglyceridemia and protection from postprandial hypertriglyceridemia. J. Biol. Chem. *269*, 23610–23616.
26. Russo, G.T., Meigs, J.B., Cupples, L.A., Demissie, S., Otvos, J.D., Wilson, P.W., Lahoz, C., Cucinotta, D., Couture, P., Mallory, T., et al. (2001). Association of the Sst-I polymorphism at the APOC3 gene locus with variations in lipid levels, lipoprotein subclass profiles and coronary heart disease risk: the Framingham offspring study. Atherosclerosis *158*, 173–181.
27. Olivieri, O., Bassi, A., Stranieri, C., Trabetti, E., Martinelli, N., Pizzolo, F., Girelli, D., Friso, S., Pignatti, P.F., and Corrocher, R. (2003). Apolipoprotein C-III, metabolic syndrome, and risk of coronary artery disease. J. Lipid Res. *44*, 2374–2381.

28. Vaessen, S.F., Schaap, F.G., Kuivenhoven, J.A., Groen, A.K., Hutten, B.A., Boekholdt, S.M., Hattori, H., Sandhu, M.S., Bingham, S.A., Luben, R., et al. (2006). Apolipoprotein A-V, triglycerides and risk of coronary artery disease: the prospective Epic-Norfolk Population Study. J. Lipid Res. *47*, 2064–2070.
29. Bi, N., Yan, S.K., Li, G.P., Yin, Z.N., and Chen, B.S. (2004). A single nucleotide polymorphism -1131T>C in the apolipoprotein A5 gene is associated with an increased risk of coronary artery disease and alters triglyceride metabolism in Chinese. Mol. Genet. Metab. *83*, 280–286.
30. Dallongeville, J., Cottel, D., Montaye, M., Codron, V., Amouyel, P., and Helbecque, N. (2006). Impact of APOA5/A4/C3 genetic polymorphisms on lipid variables and cardiovascular disease risk in French men. Int. J. Cardiol. *106*, 152–156.
31. Wijsman, E.M. (2012). The role of large pedigrees in an era of high-throughput sequencing. Hum. Genet. *131*, 1555–1563.
32. Jarvik, G.P., Brunzell, J.D., Austin, M.A., Krauss, R.M., Motulsky, A.G., and Wijsman, E. (1994). Genetic predictors of FCHL in four large pedigrees. Influence of ApoB level major locus predicted genotype and LDL subclass phenotype. Arterioscler. Thromb. *14*, 1687–1694.
33. Goldstein, J.L., Schrott, H.G., Hazzard, W.R., Bierman, E.L., and Motulsky, A.G. (1973). Hyperlipidemia in coronary heart disease. II. Genetic analysis of lipid levels in 176 families and delineation of a new inherited disorder, combined hyperlipidemia. J. Clin. Invest. *52*, 1544–1568.
34. Goldstein, J.L., Dana, S.E., Brunschede, G.Y., and Brown, M.S. (1975). Genetic heterogeneity in familial hypercholesterolemia: evidence for two different mutations affecting functions of low-density lipoprotein receptor. Proc. Natl. Acad. Sci. USA *72*, 1092–1096.
35. Breslow, J.L. (2000). Genetics of lipoprotein abnormalities associated with coronary artery disease susceptibility. Annu. Rev. Genet. *34*, 233–254.
36. Ford, E.S., Li, C., Zhao, G., Pearson, W.S., and Mokdad, A.H. (2009). Hypertriglyceridemia and its pharmacologic treatment among US adults. Arch. Intern. Med. *169*, 572–578.
37. Austin, M.A., McKnight, B., Edwards, K.L., Bradley, C.M., McNeely, M.J., Psaty, B.M., Brunzell, J.D., and Motulsky, A.G. (2000). Cardiovascular disease mortality in familial forms of hypertriglyceridemia: A 20-year prospective study. Circulation *101*, 2777–2782.
38. Chait, A., and Brunzell, J.D. (1992). Chylomicronemia syndrome. Adv. Intern. Med. *37*, 249–273.
39. Warnick, G.R. (1986). Enzymatic methods for quantification of lipoprotein lipids. Methods Enzymol. *129*, 101–123.
40. US Dept of Health and Human Services (1980). The Lipid Research Clinics' Population Studies Data Book, *Volume 1* (Washington, DC: National Institutes of Health).
41. Rosenthal, E.A., Ronald, J., Rothstein, J., Rajagopalan, R., Ranchalis, J., Wolfbauer, G., Albers, J.J., Brunzell, J.D., Motulsky, A.G., Rieder, M.J., et al. (2011). Linkage and association of phospholipid transfer protein activity to *LASS4*. J. Lipid Res. *52*, 1837–1846.
42. Wijsman, E.M., Rothstein, J.H., Igo, R.P., Jr., Brunzell, J.D., Motulsky, A.G., and Jarvik, G.P. (2010). Linkage and association analyses identify a candidate region for apoB level on chromosome 4q32.3 in FCHL families. Hum. Genet. *127*, 705–719.
43. Keating, B.J., Tischfield, S., Murray, S.S., Bhangale, T., Price, T.S., Glessner, J.T., Galver, L., Barrett, J.C., Grant, S.F.A., Farlow, D.N., et al. (2008). Concept, design and implementation of a cardiovascular gene-centric 50 k SNP array for large-scale genomic association studies. PLoS ONE *3*, e3583.
44. Ng, S.B., Turner, E.H., Robertson, P.D., Flygare, S.D., Bigham, A.W., Lee, C., Shaffer, T., Wong, M., Bhattacharjee, A., Eichler, E.E., et al. (2009). Targeted capture and massively parallel sequencing of 12 human exomes. Nature *461*, 272–276.
45. De La Vega, F.M., Dailey, D., Ziegle, J., Williams, J., Madden, D., and Gilbert, D.A. (2002). New generation pharmacogenomic tools: a SNP linkage disequilibrium Map, validated SNP assay resource, and high-throughput instrumentation system for large-scale genetic studies. Biotechniques (*Suppl*), 48–50, 52, 54.
46. Lin, C.H., Yeakley, J.M., McDaniel, T.K., and Shen, R. (2009). Medium- to high-throughput SNP genotyping using VeraCode microbeads. Methods Mol. Biol. *496*, 129–142.
47. Heath, S.C. (1997). Markov chain Monte Carlo segregation and linkage analysis for oligogenic models. Am. J. Hum. Genet. *61*, 748–760.
48. Gagnon, F., Jarvik, G.P., Badzioch, M.D., Motulsky, A.G., Brunzell, J.D., and Wijsman, E.M. (2005). Genome scan for quantitative trait loci influencing HDL levels: evidence for multilocus inheritance in familial combined hyperlipidemia. Hum. Genet. *117*, 494–505.
49. Yu, D. (2003). Testing the robustness of Markov chain Monte Carlo segregation and linkage analysis when normality assumptions are violated. PhD thesis, University of Washington, Seattle, WA.
50. Wijsman, E.M., Daw, E.W., Yu, C.E., Payami, H., Steinbart, E.J., Nochlin, D., Conlon, E.M., Bird, T.D., and Schellenberg, G.D. (2004). Evidence for a novel late-onset Alzheimer disease locus on chromosome 19p13.2. Am. J. Hum. Genet. *75*, 398–409.
51. Gagnon, F., Jarvik, G.P., Motulsky, A.G., Deeb, S.S., Brunzell, J.D., and Wijsman, E.M. (2003). Evidence of linkage of HDL level variation to *APOC3* in two samples with different ascertainment. Hum. Genet. *113*, 522–533.
52. Goldberg, I.J., Eckel, R.H., and McPherson, R. (2011). Triglycerides and heart disease: still a hypothesis? Arterioscler. Thromb. Vasc. Biol. *31*, 1716–1725.
53. Johnson, A.D., Handsaker, R.E., Pulit, S.L., Nizzari, M.M., O'Donnell, C.J., and de Bakker, P.I. (2008). SNAP: a web-based tool for identification and annotation of proxy SNPs using HapMap. Bioinformatics *24*, 2938–2939.
54. Abecasis, G.R., Altshuler, D., Auton, A., Brooks, L.D., Durbin, R.M., Gibbs, R.A., Hurles, M.E., and McVean, G.A.; 1000 Genomes Project Consortium (2010). A map of human genome variation from population-scale sequencing. Nature *467*, 1061–1073.
55. Sherry, S.T., Ward, M.H., Kholodov, M., Baker, J., Phan, L., Smigielski, E.M., and Sirotkin, K. (2001). dbSNP: the NCBI database of genetic variation. Nucleic Acids Res. *29*, 308–311.
56. Cooper, G.M., Stone, E.A., Asimenos, G., Green, E.D., Batzoglou, S., and Sidow, A.; NISC Comparative Sequencing Program (2005). Distribution and intensity of constraint in mammalian genomic sequence. Genome Res. *15*, 901–913.
57. Almasy, L., and Blangero, J. (1998). Multipoint quantitative-trait linkage analysis in general pedigrees. Am. J. Hum. Genet. *62*, 1198–1211.
58. Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M., and Haussler, D. (2002). The human genome browser at UCSC. Genome Res. *12*, 996–1006.

59. Davydov, E.V., Goode, D.L., Sirota, M., Cooper, G.M., Sidow, A., and Batzoglou, S. (2010). Identifying a high fraction of the human genome to be under selective constraint using GERP++. PLoS Comput. Biol. *6*, e1001025.
60. Adzhubei, I.A., Schmidt, S., Peshkin, L., Ramensky, V.E., Gerasimova, A., Bork, P., Kondrashov, A.S., and Sunyaev, S.R. (2010). A method and server for predicting damaging missense mutations. Nat. Methods *7*, 248–249.
61. Pagliarini, D.J., Calvo, S.E., Chang, B., Sheth, S.A., Vafai, S.B., Ong, S.E., Walford, G.A., Sugiana, C., Boneh, A., Chen, W.K., et al. (2008). A mitochondrial protein compendium elucidates complex I disease biology. Cell *134*, 112–123.
62. Pontius, J.U., Wagner, L., and Schuler, G.D. (2003). UniGene: a unified view of the transcriptome. In The NCBI Handbook, J. McEntyre and J. Ostell, eds. (Bethesda, MD: National Center for Biotechnology Information).
63. Siepel, A., Bejerano, G., Pedersen, J.S., Hinrichs, A.S., Hou, M., Rosenbloom, K., Clawson, H., Spieth, J., Hillier, L.W., Richards, S., et al. (2005). Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res. *15*, 1034–1050.
64. Grantham, R. (1974). Amino acid difference formula to help explain protein evolution. Science *185*, 862–864.
65. Kim, P.S., and Baldwin, R.L. (1982). Specific intermediates in the folding reactions of small proteins and the mechanism of protein folding. Annu. Rev. Biochem. *51*, 459–489.
66. Blackwood, E.M., and Eisenman, R.N. (1991). Max: a helix-loop-helix zipper protein that forms a sequence-specific DNA-binding complex with Myc. Science *251*, 1211–1217.
67. Kretzner, L., Blackwood, E.M., and Eisenman, R.N. (1992). Myc and Max proteins possess distinct transcriptional activities. Nature *359*, 426–429.

Original Investigation

Association of Aspirin and NSAID Use With Risk of Colorectal Cancer According to Genetic Variants

Hongmei Nan, MD, PhD; Carolyn M. Hutter, MS, PhD; Yi Lin, MS; Eric J. Jacobs, PhD, MS; Cornelia M. Ulrich, PhD; Emily White, PhD; John A. Baron, MD; Sonja I. Berndt, PharmD, PhD; Hermann Brenner, MD, MPH; Katja Butterbach, PhD; Bette J. Caan, DrPH; Peter T. Campbell, PhD; Christopher S. Carlson, PhD; Graham Casey, PhD; Jenny Chang-Claude, PhD; Stephen J. Chanock, MD; Michelle Cotterchio, PhD; David Duggan, PhD; Jane C. Figueiredo, PhD; Charles S. Fuchs, MD, MPH; Edward L. Giovannucci, MD; Jian Gong, PhD; Robert W. Haile, DrPH; Tabitha A. Harrison, MPH; Richard B. Hayes, DDS, PhD; Michael Hoffmeister, PhD; John L. Hopper, PhD; Thomas J. Hudson, MD; Mark A. Jenkins, PhD; Shuo Jiao, PhD; Noralane M. Lindor, MD; Mathieu Lemire, PhD; Loic Le Marchand, MD, PhD; Polly A. Newcomb, PhD, MPH; Shuji Ogino, MD, PhD; Bethann M. Pflugeisen, PhD; John D. Potter, MD, PhD; Conghui Qu, MS; Stephanie A. Rosse, PhD; Anja Rudolph, PhD; Robert E. Schoen, MD, MPH; Fredrick R. Schumacher, PhD; Daniela Seminara, PhD, MPH; Martha L. Slattery, PhD; Stephen N. Thibodeau, PhD; Fridtjof Thomas, PhD; Mark Thornquist, PhD; Greg S. Warnick; Brent W. Zanke, MD, PhD, FRCPC; W. James Gauderman, PhD; Ulrike Peters, PhD, MPH; Li Hsu, PhD; Andrew T. Chan, MD, MPH; for the CCFR and GECCO

IMPORTANCE Use of aspirin and other nonsteroidal anti-inflammatory drugs (NSAIDs) is associated with lower risk of colorectal cancer.

OBJECTIVE To identify common genetic markers that may confer differential benefit from aspirin or NSAID chemoprevention, we tested gene × environment interactions between regular use of aspirin and/or NSAIDs and single-nucleotide polymorphisms (SNPs) in relation to risk of colorectal cancer.

DESIGN, SETTING, AND PARTICIPANTS Case-control study using data from 5 case-control and 5 cohort studies initiated between 1976 and 2003 across the United States, Canada, Australia, and Germany and including colorectal cancer cases (n=8634) and matched controls (n=8553) ascertained between 1976 and 2011. Participants were all of European descent.

EXPOSURES Genome-wide SNP data and information on regular use of aspirin and/or NSAIDs and other risk factors.

MAIN OUTCOMES AND MEASURES Colorectal cancer.

RESULTS Regular use of aspirin and/or NSAIDs was associated with lower risk of colorectal cancer (prevalence, 28% vs 38%; odds ratio [OR], 0.69 [95% CI, 0.64-0.74]; $P = 6.2 \times 10^{-28}$) compared with nonregular use. In the conventional logistic regression analysis, the SNP rs2965667 at chromosome 12p12.3 near the *MGST1* gene showed a genome-wide significant interaction with aspirin and/or NSAID use ($P = 4.6 \times 10^{-9}$ for interaction). Aspirin and/or NSAID use was associated with a lower risk of colorectal cancer among individuals with rs2965667-TT genotype (prevalence, 28% vs 38%; OR, 0.66 [95% CI, 0.61-0.70]; $P = 7.7 \times 10^{-33}$) but with a higher risk among those with rare (4%) TA or AA genotypes (prevalence, 35% vs 29%; OR, 1.89 [95% CI, 1.27-2.81]; $P = .002$). In case-only interaction analysis, the SNP rs16973225 at chromosome 15q25.2 near the *IL16* gene showed a genome-wide significant interaction with use of aspirin and/or NSAIDs ($P = 8.2 \times 10^{-9}$ for interaction). Regular use was associated with a lower risk of colorectal cancer among individuals with rs16973225-AA genotype (prevalence, 28% vs 38%; OR, 0.66 [95% CI, 0.62-0.71]; $P = 1.9 \times 10^{-30}$) but was not associated with risk of colorectal cancer among those with less common (9%) AC or CC genotypes (prevalence, 36% vs 39%; OR, 0.97 [95% CI, 0.78-1.20]; $P = .76$).

CONCLUSIONS AND RELEVANCE In this genome-wide investigation of gene × environment interactions, use of aspirin and/or NSAIDs was associated with lower risk of colorectal cancer, and this association differed according to genetic variation at 2 SNPs at chromosomes 12 and 15. Validation of these findings in additional populations may facilitate targeted colorectal cancer prevention strategies.

JAMA. 2015;313(11):1133-1142. doi:10.1001/jama.2015.1815
Corrected on March 25, 2015.

Editorial page 1111

Supplemental content at jama.com

Author Affiliations: Author affiliations are listed at the end of this article.

Corresponding Authors: Li Hsu, PhD, Fred Hutchinson Cancer Research Center, 1100 Fairview Avenue North, M2-B500, Seattle, WA 98109 (lih@fredhutch.org); Andrew T. Chan, MD, MPH, Division of Gastroenterology, Massachusetts General Hospital, GRJ-825C, Boston, MA 02114 (achan@mgh.harvard.edu).

Considerable evidence demonstrates that use of aspirin and other nonsteroidal anti-inflammatory drugs (NSAIDs) is associated with lower risk of colorectal neoplasms.[1-5] However, the mechanisms behind this association are not well understood. Routine use of aspirin, NSAIDs, or both for chemoprevention of cancer is not currently recommended because of uncertainty about risk-benefit profile. Hence, understanding the interrelationship between genetic markers and use of aspirin and NSAIDs, also known as gene × environment interactions, can help to identify population subgroups defined by genetic background that may preferentially benefit from chemopreventive use of these agents and offer novel insights into underlying mechanisms of carcinogenesis.

CCFR *Colon Cancer Family Registry*

COX-2 *cyclooxygenase 2*

GECCO *Genetics and Epidemiology of Colorectal Cancer Consortium*

GWAS *genome-wide association study*

MAF *minor allele frequency*

NSAID *nonsteroidal anti-inflammatory drug*

PGE_2 *proinflammatory prostaglandin E_2*

PTGS2 *prostaglandin-endoperoxide synthase 2*

SNP *single-nucleotide polymorphism*

Previous genetic studies have examined the association of aspirin, NSAIDs, or both with colorectal cancer according to a limited number of candidate genes or pathways.[6-10] Thus, to comprehensively identify common genetic markers that characterize individuals who may obtain differential benefit from aspirin and NSAIDs, we conducted a discovery-based, genome-wide analysis of gene × environment interactions between regular use of aspirin, NSAIDs, or both and single-nucleotide polymorphisms (SNPs) in relation to risk of colorectal cancer.

Methods

Study Population and Harmonization of Environmental Data

We included individual-level data pooled from a case-control study from the Colon Cancer Family Registry (CCFR) and 9 studies from the Genetics and Epidemiology of Colorectal Cancer Consortium (GECCO) that were initiated between 1976 and 2003 and that enrolled cases of colorectal cancer diagnosed between 1976 and 2011 and matched controls across the United States, Canada, Australia, and Germany (**Table 1**). The cohorts are described in the eAppendix in the Supplement. All cases were defined as invasive colorectal adenocarcinoma and confirmed by medical record, pathology report, or death certificate. For prospective cohorts, nested case-control sets were constructed by fixing the cohort at a point at which risk set sampling was used to select cases and controls. For other case-control studies, population-based controls were used. For all studies, controls were matched on age, sex, and race/ethnicity; for some studies, controls were also matched on additional factors, such as enrollment date and trial group.

Study-specific eligibility and our multistep data harmonization procedure are described in the eAppendix in the Supplement. Briefly, within each study, all exposure information, including use of aspirin, NSAIDs, or both, was collected by in-person interviews, structured questionnaires, or both with the reference time for cohort studies as the time of enrollment (Women's Health Initiative [WHI]; Prostate, Lung, Colorectal and Ovarian Cancer Screening Trial [PLCO]; and Vitamins and Lifestyle [VITAL]) or blood draw (Health Professionals Follow-up Study [HPFS] and Nurses' Health Study [NHS]). Individuals with missing data on use of aspirin and NSAIDs were excluded. The precise definition of regular use of aspirin, NSAIDs, or both, which was determined individually by each study cohort, is provided in Table 1.

All participants provided written or oral informed consent, and studies were reviewed and approved by their respective institutional review boards or ethics committees.

Statistical Methods

A detailed description for genotyping, quality assurance and quality control, and imputation is provided in the Supplement. Mean sample and SNP call rates, and concordance rates for blinded duplicates, are listed in eTable 1 in the Supplement. In brief, genotyped SNPs were excluded based on call rate (<98%), lack of Hardy-Weinberg equilibrium (HWE) in controls ($P < 1 \times 10^{-4}$), and minor allele frequency (MAF) (MAF <5% for WHI Set 1, Diet, Activity and Lifestyle Study [DALS] Set 1, and Ontario Familial Colorectal Cancer Registry [OFCCR]; MAF <5/No. of samples for each other study). Because imputation of genotypes is standard practice in genetic association analysis, all autosomal SNPs of each study were imputed to the CEPH collection (CEU) population in HapMap II using IMPUTE (CCFR), BEAGLE (OFCCR), and MACH (all other studies).

After imputation and quality-control analyses, a total of approximately 2.7 million SNPs were used in the analysis. To reduce heterogeneity, all analyses were restricted to samples self-reported as of European descent and clustering with Utah residents with Northern/Western European ancestry from the CEU population in principal component analysis, including the HapMap II populations as reference.

Statistical analyses were conducted centrally on individual-level data. We adjusted for age at reference time, sex, center, and racial composition using the first 3 principal components from EIGENSTRAT to account for population substructure. Each directly genotyped SNP was coded as 0, 1, or 2 copies of the variant allele. For imputed SNPs, we used the expected number of copies of the variant allele, which provides unbiased test statistics.[11] Both genotyped and imputed SNPs were examined as continuous variables (ie, assuming log-additive effects).

We analyzed each study separately using logistic regression models and combined study-specific results using fixed effect to obtain summary odds ratios (ORs) and 95% CIs. We calculated *P* values for heterogeneity using the Cochran Q test.[12] Fixed-effect meta-analysis is routinely used in genome-wide association studies (GWAS) because it is the most powerful approach for identifying disease-associated variants.[13,14] Furthermore, in our study fixed effect was more appropriate than

Table 1. Descriptive Characteristics of Study Populations

Study	Design	Country	Years		No.		Age, Mean (Range), y	Female, No. (%)	Covariates Used in Base Model Analysis[a]	Definition of Regular Use of Aspirin and/or NSAIDs[b]
			Inception or Recruitment	Diagnosis	Cases	Controls				
CCFR	Case-control	United States, Canada, Australia	1998-2006	1998-2006[c]	1163	978	54.3 (17-81)	1067 (49.8)	Age, sex, 3 principal components, center	At least twice a wk for >1 mo
DACHS	Case-control	Germany	2003-2010	2003-2010	2339	2180	68.7 (33-99)	1801 (39.9)	Age, sex, 3 principal components	At least 1 time/mo for ≥1 y
DALS	Case-control	United States	1991-1994	1991-1994	1115	1173	63.8 (28-79)	1027 (44.9)	Age, sex, 3 principal components, center	At least 3 times/wk for ≥1 mo
HPFS	Cohort	United States	1986	1986-2008	403	401	65.2 (48-83)	0	Age, 3 principal components	Currently taking at least 2 times/wk
NHS	Cohort	United States	1976	1976-2008	553	955	59.7 (44-69)	1508 (100)	Age, 3 principal components	On average ≥5 d/mo
OFCCR	Case-control	Canada	2000-2006	1998-2003	553	519	62.1 (29-77)	577 (53.8)	Age, sex, 3 principal components	At least twice/wk for >1 mo
PMH-CCFR	Case-control	United States	1998-2003	1998-2002[d]	280	122	62.8 (48-73)	402 (100)	Age, 3 principal components	At least twice/wk for >1 mo
PLCO	Cohort	United States	1993-2001	1994-2009	485	415	63.6 (55-75)	382 (42.4)	Age, sex, 3 principal components, center	At least twice/wk in the last 12 mo
VITAL	Cohort	United States	2000-2002	2000-2009	277	279	66.5 (50-76)	268 (48.2)	Age, sex, 3 principal components	≥4 d/wk for 1 y
WHI	Cohort	United States	1993-1998	1993-2011	1466	1531	66.3 (50-79)	2997 (100)	Age, 3 principal components, region	At least once/wk for at least the last 2 wk

Abbreviations: CCFR, Colon Cancer Family Registry; DACHS, Darmkrebs: Chancen der Verhütung durch Screening Study; DALS, Diet, Activity and Lifestyle Study; HPFS, Health Professionals Follow-up Study; NHS, Nurses' Health Study; NSAID, nonsteroidal anti-inflammatory drug; OFCCR, Ontario Familial Colorectal Cancer Registry; PMH-CCFR, Postmenopausal Hormone Study–Colon Cancer Family Registry; PLCO, Prostate, Lung, Colorectal and Ovarian Cancer Screening Trial; VITAL, Vitamins and Lifestyle; WHI, Women's Health Initiative.

[a] In the multivariable-adjusted model, in addition to the covariates adjusted in the base model, those colorectal cancer-related factors, including smoking status (never, former, or current smoker), body mass index, alcohol consumption, and red meat consumption, are also adjusted.

[b] Includes regular use of aspirin only, NSAIDs only, or both aspirin and NSAIDs.

[c] All cases diagnosed between 1998-2006, except 2 cases for which the year of diagnosis is 1985 and 1995.

[d] All cases diagnosed between 1998-2002, except 4 cases for which the year of diagnosis is 1983, 1984, 1990, 1991.

random effects, since the Q-Q plots and the *P* value distributions indicated minimal heterogeneity across studies. Moreover, the effects may not fit a Gaussian distribution as required by the random-effects model, and the limited number of included studies may lead to an imprecise estimate of heterogeneity.[15]

To test for gene × environment interactions between SNPs and the regular use of aspirin, NSAIDs, or both (including use of aspirin only, NSAIDs only, or both aspirin and NSAIDs) or the regular use of aspirin only, we used conventional case-control logistic regression and case-only interaction analyses. Equations for the models used in the interaction analyses are provided in the eAppendix in the Supplement. We examined genome-wide correlations between SNPs and use of aspirin, NSAIDs, or both using linear regression analysis and did not observe deviation from independence. For all genome-wide gene × environment interaction analyses, $P < 5.0 \times 10^{-8}$ (2-sided), which yields a genome-wide significance level of .05, was considered statistically significant.

As described in the eAppendix in the Supplement, for each SNP showing gene × environment interaction with use of aspirin, NSAIDs, or both, we estimated the association of such use with colorectal cancer risk stratified by SNP genotypes, as well as associations in strata defined by SNP and use of aspirin, NSAIDs, or both with 1 common reference group. We also estimated absolute risks associated with use of aspirin, NSAIDs, or both among individuals defined by specific genotypes based on Surveillance, Epidemiology, and End Results age-adjusted colorectal cancer incidence rates (eAppendix in the Supplement).

All analyses were conducted using R 3.1.2 (R Foundation for Statistical Computing [http://www-r-project.org]).

Results

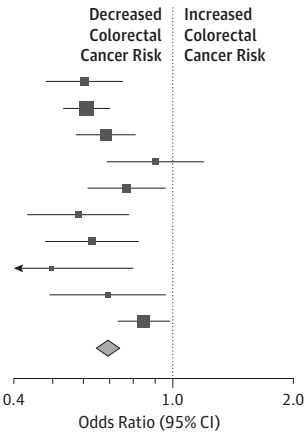
The characteristics of the 8634 colorectal cancer cases and 8553 controls of European descent within each cohort from the CCFR and GECCO are provided in Table 1. As shown in the **Figure**, compared with nonregular use, regular use of aspirin, NSAIDs, or both (prevalence, 28% vs 38%; OR, 0.69 [95% CI, 0.64-0.74]; $P = 6.2 \times 10^{-28}$; $P = .02$ for heterogeneity) or aspirin only

Figure. Main Associations of Regular Use of Aspirin, NSAIDs, or Both and Aspirin Only With the Risk of Colorectal Cancer

A Aspirin, NSAIDs, or both

Study	Colorectal Cancer Diagnosis No./Total (%)		Odds Ratio (95% CI)
	Cases	Controls	
CCFR	204/1163 (17.5)	297/978 (30.4)	0.60 (0.48–0.75)
DACHS	544/2339 (23.3)	729/2180 (33.4)	0.61 (0.53–0.70)
DALS	370/1115 (33.2)	494/1173 (42.1)	0.68 (0.57–0.81)
HPFS	184/403 (45.7)	192/401 (47.9)	0.90 (0.68–1.20)
NHS	172/553 (31.1)	362/955 (37.9)	0.77 (0.61–0.96)
OFCCR	101/553 (18.3)	159/519 (30.6)	0.58 (0.43–0.78)
PLCO	205/485 (42.3)	224/415 (54.0)	0.63 (0.48–0.82)
PMH–CCFR	62/280 (22.1)	43/122 (35.2)	0.50 (0.31–0.80)
VITAL	120/277 (43.3)	147/279 (52.7)	0.69 (0.49–0.96)
WHI	493/1466 (33.6)	574/1531 (37.5)	0.85 (0.73–0.98)
Overall	2455/8634 (28.4)	3221/8553 (37.7)	0.69 (0.64–0.74)

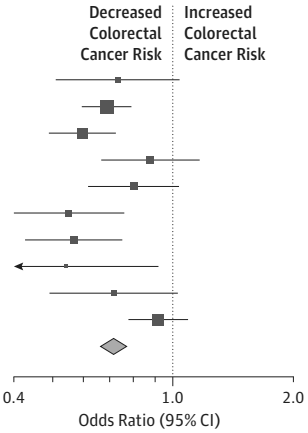
$P = 6.2 \times 10^{-28}$; $P = .02$ for heterogeneity



B Aspirin only

Study	Colorectal Cancer Diagnosis No./Total (%)		Odds Ratio (95% CI)
	Cases	Controls	
CCFR	141/260 (54.2)	236/360 (65.6)	0.73 (0.51–1.04)
DACHS	470/2329 (20.2)	597/2179 (27.4)	0.68 (0.59–0.79)
DALS	234/1110 (21.1)	360/1167 (30.8)	0.59 (0.49–0.72)
HPFS	157/403 (39.0)	168/401 (41.9)	0.88 (0.66–1.17)
NHS	107/553 (19.3)	230/955 (24.1)	0.80 (0.61–1.04)
OFCCR	77/547 (14.1)	136/512 (26.6)	0.55 (0.40–0.76)
PLCO	151/484 (31.2)	183/415 (44.1)	0.56 (0.43–0.75)
PMH–CCFR	46/274 (16.8)	30/121 (24.8)	0.54 (0.32–0.92)
VITAL	103/230 (44.8)	126/236 (53.4)	0.71 (0.49–1.03)
WHI	329/1466 (22.4)	367/1531 (24.0)	0.92 (0.77–1.09)
Overall	1815/7656 (23.7)	2433/7877 (30.9)	0.71 (0.66–0.77)

$P = 4.96 \times 10^{-19}$; $P = .01$ for heterogeneity



The size of the data markers is proportional to the precision of the estimate, which is the inverse of the variance. NSAID indicates nonsteroidal anti-inflammatory drug. For expansions of study names, see Table 1.

(prevalence, 24% vs 31%; OR, 0.71 [95% CI, 0.66-0.77]; $P = 5.0 \times 10^{-19}$; $P = .01$ for heterogeneity) was associated with lower risk of colorectal cancer.

For the conventional logistic regression interaction analysis between each SNP and aspirin and/or NSAID use, the *P* values are shown in the Manhattan plot and Q-Q plot (eFigure 1 in the Supplement). At chromosome 12p12.3, we observed SNP rs2965667 (MAF = 1.7%) showing a genome-wide significant interaction with regular use of aspirin, NSAIDs, or both ($P = 4.6 \times 10^{-9}$ for interaction). The SNP rs10505806 (MAF = 3.8%), which had the second-lowest *P* value, was also found in the same locus, but it did not reach genome-wide significant interaction ($P = 5.5 \times 10^{-8}$ for interaction). These 2 top SNPs (rs2965667 and rs10505806) were highly correlated ($D' = 1.0$ and $r^2 = 0.74$ in HapMap CEU). In stratified analysis, compared with nonregular use, regular use of aspirin, NSAIDs, or both was statistically significantly associated with lower risk of colorectal cancer among individuals with rs2965667-TT genotype (prevalence, 28% vs 38%; OR, 0.66 [95% CI, 0.61-0.70]; $P = 7.7 \times 10^{-33}$), which comprised 96% (n = 16 465) of the population. In contrast, a higher risk was observed among the 4% (n = 722) of the population with TA or AA genotypes (prevalence, 35% vs 29%; OR, 1.89 [95% CI, 1.27-2.81]; $P = .002$).

As expected, stratified results for the highly correlated rs10505806 were similar to those for rs2965667. Compared with nonregular use, regular use of aspirin, NSAIDs, or both was statistically significantly associated with lower risk of colorectal cancer among individuals with rs10505806-AA genotype (prevalence, 28% vs 38%; OR, 0.66 [95% CI, 0.61-0.70]; $P = 8.7 \times 10^{-33}$), which comprised 95% (n = 16 328) of the population. In contrast, a higher risk was observed among the 5% (n = 859) of the population with AT or TT genotypes (prevalence, 35% vs 31%; OR, 1.56 [95% CI, 1.12-2.16]; $P = .008$) (**Table 2** and eFigure 2 in the Supplement). The SNP rs2965667 also appeared as the SNP with the lowest *P* value in the exploratory analyses of aspirin only, but it did not reach genome-wide significant interaction ($P = 8.0 \times 10^{-7}$ for interaction; $P = .35$ for heterogeneity) (eTable 2 in the Supplement).

Both of these 2 highly correlated SNPs (rs2965667 and rs10505806) were imputed across all studies (100% of study

Table 2. Risk for Colorectal Cancer According to Regular Use of Aspirin and/or NSAIDs, Stratified by the Genotypes of rs2965667, rs10505806, and rs16973225[a]

SNP/Genotype	Use of Aspirin and/or NSAIDs		P Value
	Nonregular	Regular[b]	
rs2965667[c]			
TT			
Participants, No.			
Cases	5933	2325	
Controls	5088	3119	
OR (95% CI)			
Base model[d]	1 [Reference]	0.66 (0.61-0.70)	7.7×10^{-33}
Multivariable-adjusted model[e]	1 [Reference]	0.63 (0.59-0.68)	2.3×10^{-35}
TA or AA			
Participants, No.			
Cases	246	130	
Controls	244	102	
OR (95% CI)			
Base model[d]	1 [Reference]	1.89 (1.27-2.81)	.002
Multivariable-adjusted model[e]	1 [Reference]	1.76 (1.16-2.66)	.008
P value for interaction[f]		4.6×10^{-9}	
rs10505806[c]			
AA			
Participants, No.			
Cases	5896	2301	
Controls	5039	3092	
OR (95% CI)			
Base model[d]	1 [Reference]	0.66 (0.61-0.70)	8.7×10^{-33}
Multivariable-adjusted model[e]	1 [Reference]	0.63 (0.59-0.68)	4.2×10^{-35}
AT or TT			
Participants, No.			
Cases	283	154	
Controls	293	129	
OR (95% CI)			
Base model[d]	1 [Reference]	1.56 (1.12-2.16)	.008
Multivariable-adjusted model[e]	1 [Reference]	1.42 (1.01-2.00)	.045
P value for interaction[f]		5.5×10^{-8}	

(continued)

samples), with a mean imputation R^2 of 0.7 for rs2965667 and 0.8 for rs10505806 (eTable 3 in the Supplement). To further validate accuracy of imputation, we conducted direct genotyping of rs10505806 in participants enrolled in the NHS (553 cases and 955 controls) and the HPFS (403 cases and 401 controls).

The overall concordance of the SNP rs10505806 between imputed vs genotyped data was high (Pearson correlation coefficient *r* of 0.89). Among the total 956 cases and 1356 controls within NHS and HPFS whom we also directly genotyped rs10505806, we compared the gene × environment interaction statistical effect using direct genotype data with the imputed data. We confirmed no material difference in interaction estimates ($P = .50$ for heterogeneity) between imputed data (OR, 2.57 [95% CI, 1.02-6.43]; $P = .045$ for interaction) and directly genotyped data (OR, 2.19 [95% CI, 1.04-4.59]; $P = .04$ for interaction).

In case-only interaction analysis, SNP rs16973225 at chromosome 15q25.2 showed a genome-wide significant interaction with regular use of aspirin, NSAIDs, or both ($P = 8.2 \times 10^{-9}$ for interaction). In the stratified analysis, compared with nonregular use, regular use of aspirin, NSAIDs, or both was statistically significantly associated with lower risk of colorectal

Table 2. Risk for Colorectal Cancer According to Regular Use of Aspirin and/or NSAIDs, Stratified by the Genotypes of rs2965667, rs10505806, and rs16973225[a] (continued)

SNP/Genotype	Use of Aspirin and/or NSAIDs		*P* Value
	Nonregular	Regular[b]	
rs16973225[g]			
AA			
Participants, No.			
Cases	5686	2181	
Controls	4840	2909	
OR (95% CI)			
Base model[d]	1 [Reference]	0.66 (0.62-0.71)	1.9×10^{-30}
Multivariable-adjusted model[e]	1 [Reference]	0.63 (0.59-0.68)	3.5×10^{-33}
AC or CC			
Participants, No.			
Cases	491	274	
Controls	492	311	
OR (95% CI)			
Base model[d]	1 [Reference]	0.97 (0.78-1.20)	.76
Multivariable-adjusted model[e]	1 [Reference]	0.93 (0.75-1.17)	.55
P value for interaction[f]		8.2×10^{-9}	

Abbreviations: NSAID, nonsteroidal anti-inflammatory drug; OR, odds ratio; SNP, single-nucleotide polymorphism.

[a] The numbers of cases and controls were from the base model. For the SNP rs16973225, the total sample size is slightly smaller than in Table 1 because of missing genotype (n = 3).

[b] Regular use of aspirin only, NSAIDs only, or both aspirin and NSAIDs.

[c] SNPs rs2965667 and rs10505806 were identified from conventional logistic regression analysis.

[d] Odds ratios in base models are adjusted for age at the reference time, sex, center, and the first 3 principal components from EIGENSTRAT.

[e] Odds ratios in multivariable-adjusted models are adjusted for age at the reference time, sex, center, the first 3 principal components from EIGENSTRAT, smoking status (never, former, or current smoker), body mass index, alcohol consumption, and red meat consumption.

[f] *P* values for interaction were calculated after adjusting for age at the reference time, sex, center, and the first 3 principal components from EIGENSTRAT.

[g] SNP rs16973225 was identified from case-only interaction analysis.

cancer among individuals with rs16973225-AA genotype (prevalence, 28% vs 38%; OR, 0.66 [95% CI, 0.62-0.71]; $P = 1.9 \times 10^{-30}$), which comprised 91% (n = 15 616) of the population, but was not associated with risk of colorectal cancer among those with AC or CC genotypes (prevalence, 36% vs 39%; OR, 0.97 [95% CI, 0.78-1.20]; $P = .76$) (Table 2 and eFigure 2 in the Supplement), which comprised 9% (n = 1568) of the population.

The SNP rs16973225 was directly genotyped in 9 of 15 study sets and was imputed with high quality ($R^2 = 0.9$) in the remaining 6 study sets (38% of study samples) (eTable 3 in the Supplement). To validate imputation of rs16973225, we compared the gene × environment interaction statistical effect with colorectal cancer between imputed vs genotyped study sets in case-only interaction analysis. We found that the interaction statistical effect size was not different ($P = .73$ for heterogeneity) within cohorts based on imputed data (OR, 1.68 [95% CI, 1.30-2.17]; $P = 4.7 \times 10^{-5}$ for interaction) compared with cohorts based on directly genotyped data (OR, 1.59 [95% CI, 1.28-1.97]; $P = 4.2 \times 10^{-5}$ for interaction). In the case-only analysis of aspirin only, we did not observe genome-wide significant interactions.

The SNP rs2965667 showing a genome-wide significant interaction with use of aspirin, NSAIDs, or both in conventional logistic regression case-control analysis also appeared as a notable variant in case-only interaction analysis, although it did not achieve a genome-wide significance level ($P = 7.5 \times 10^{-8}$ for interaction). Similarly, the SNP rs16973225 reaching a genome-wide significant interaction with use of aspirin, NSAIDs, or both in case-only interaction analysis also showed evidence for gene × environment interaction in conventional logistic regression analysis ($P = 2.2 \times 10^{-4}$ for interaction).

The results for the 3 SNPs showing gene × environment interaction (rs2965667, rs10505806, and rs16973225) did not materially change after adjusting for additional colorectal cancer risk factors, including smoking status, body mass index, alcohol consumption, and red meat consumption (Table 2 and eTable 4 in the Supplement). For these 3 SNPs, we report the ORs for use of aspirin, NSAIDs, or both across genotypes corresponding to 0, 1, or 2 copies of the variant allele (eTable 5 in the Supplement) and the ORs for each SNP by strata of use of aspirin, NSAIDs, or both with 1 common reference group (eTable 6 in the Supplement), to fully describe the interaction.

We estimated absolute risks associated with use of aspirin, NSAIDs, or both among individuals with specific genotypes defined by each of these 3 SNPs. Compared with nonuse of aspirin, NSAIDs, or both, regular use was associated with 16.6 fewer colorectal cancer cases per 100 000 individuals with the rs2965667-TT genotype per year; 16.7 fewer colorectal cancer cases per 100 000 individuals with the rs10505806-AA genotype per year; and 16.8 fewer colorectal cancer cases per 100 000 individuals with the rs16973225-AA genotype per year. In contrast, regular use of aspirin, NSAIDs, or both was associated with 34.7 additional colorectal cancer cases per 100 000 individuals with rs2965667-TA or -AA genotypes per year; 21.1 additional colorectal cancer cases per 100 000 individuals with rs10505806-AT or -TT genotypes per year; and only 1.5 fewer colorectal cancer cases per 100 000 with rs16973225-AC or -CC genotypes per year.

Discussion

Consistent with the preponderance of experimental, epidemiologic, and clinical trial evidence, we found that use of aspirin, NSAIDs, or both was associated with overall lower risk of colorectal cancer in this large genome-wide investigation of gene × environment interaction, which included 8634 colorectal cancer cases and 8553 controls. However, we identified that use of aspirin, NSAIDs, or both was differentially associated with colorectal cancer risk according to genetic variation at 2 highly correlated SNPs at chromosome 12p12.3 (rs2965667 and rs10505806) using a conventional logistic regression analysis.

These SNPs are 927 kb to 971 kb downstream from microsomal glutathione S-transferase 1 (*MGST1* [NCBI Entrez Gene 4257]) (eFigure 3 in the Supplement), a member of the superfamily of membrane-associated proteins in eicosanoid and glutathione metabolism (MAPEG). *MGST1* has high sequence homology to prostaglandin E synthase (*MGST1L1* [NCBI Entrez Gene 9536]), another homologue of the MAPEG family that shares 38% of its DNA sequences with *MGST1*.[16] MGST1 and MGST1L1 are up-regulated in several cancers, including colorectal cancer.[17,18] MGST1L1 is coexpressed and functionally coupled to prostaglandin-endoperoxide synthase 2 (PTGS2; also known as cyclooxygenase 2 [COX-2]), and the combined activity of MGST1L1 and COX-2 increases production of proinflammatory prostaglandin E_2 (PGE_2), which promotes carcinogenesis through several mechanisms, including stimulation of *WNT* signaling, an essential oncogenic pathway of colorectal cancer.[19-22] An in vitro experiment has demonstrated that NSAIDs can inhibit expression of MGST1L1 and COX-2, thereby blocking COX-2–mediated synthesis of PGE_2 in human colon carcinoma cells.[23]

Taken together, both *MGST1L1* and the closely related gene *MGST1* may influence NSAID-mediated inhibition of colorectal carcinogenesis partially through involvement in the PGE_2-induced *WNT* signaling pathway. This finding is consistent with strong biological evidence linking genes in *WNT* signaling; use of aspirin, NSAIDs, or both; and colorectal cancer.[24,25]

Another candidate gene in this region is LIM domain only 3 (*LMO3* [NCBI Entrez Gene 55885], a known oncogene located about 686 kb upstream from rs2965667 (eFigure 3 in the Supplement). Altered expression of *LMO3* may contribute to the development of several cancers, such as neuroblastoma and lung cancer.[26,27]

The SNP rs2965667 is also located about 970 kb upstream from phosphatidylinositol-4-phosphate 3-kinase, catalytic subunit type 2 gamma (*PIK3C2G* [NCBI Entrez Gene 5288]) (eFigure 3 in the Supplement). The protein encoded by the *PIK3C2G* gene belongs to the phosphatidylinositol-4,5-bisphosphonate 3-kinase (PI3K) family, which plays a critical role in cancer.[28] Experimental evidence suggests that activation of PI3K signaling enhances production of COX-2 and PGE_2, which results in inhibition of apoptosis in colon cancer cell lines that can be restored with NSAID-mediated blockade of PI3K.[29]

Moreover, our previous study found that regular use of aspirin after diagnosis was associated with longer survival among the 15% to 30% of patients with colorectal cancer and with a mutation in phosphatidylinositol-4,5-bisphosphate 3-kinase, catalytic subunit alpha (*PIK3CA* [NCBI Entrez Gene 5290]), one of the PI3K family genes.[30] Markedly improved survival associated with aspirin according to *PIK3CA* status was also found in an analysis within a separate clinical trial cohort.[31] Further investigations for the joint effect of these genes would be helpful to better understand the underlying molecular mechanisms of aspirin, NSAIDs, and colorectal cancer.

In the case-only interaction analysis, another SNP, rs16973225 at chromosome 15q25.2, was identified with genome-wide significant association. This SNP is about 625 kb upstream of interleukin 16 (*IL16* [NCBI Entrez Gene 3603) (eFigure 4 in the Supplement). As a multifunctional cytokine, IL16 plays a critical role in proinflammatory processes, including inflammatory bowel disease, *Clostridium difficile*–associated colitis, and many cancers, including colorectal.[32-34] Moreover, IL16 may stimulate monocyte induction of proinflammatory cytokines associated with tumorigenesis, including IL6 and tumor necrosis factor α,[35,36] induction of COX-2 expression, and activation of *WNT* signaling.[36] This evidence suggests the possibility that polymorphisms in or near the *IL16* gene may regulate the production of inflammatory cytokines that modify the chemopreventive effect of aspirin or NSAIDs on colorectal cancer. It is plausible that those GWAS-identified promising loci outside of known coding regions affect more distant genes rather than the closest gene, since GWAS loci may be enhancers that can influence gene expression over a span of several hundred kilobases.[37]

Our study has several strengths. First, our large sample size facilitated detection of genome-wide gene × environment interactions, even using a conventional logistic regression or case-only interaction analysis and accounting for the stringent threshold for statistical significance. Second, we identified variants near genes possessing high functional plausibility given their critical roles in inflammation and prostaglandin synthesis, which have been mechanistically

linked to use of aspirin or NSAIDs and colorectal carcinogenesis.

We acknowledge some limitations. First, heterogeneity exists in the definition of regular use of aspirin, NSAIDs, or both and the range of exposure periods encompassed by each study. However, we used a standardized harmonization process on a range of environmental variables, including use of aspirin, NSAIDs, or both across 10 cohort and case-control studies. The forest plots (Figure) show the consistency of the association between use of aspirin, NSAIDs, or both and colorectal cancer on a per-study level, and the pooled risk estimate (ie, OR) is remarkably similar to those from prior studies.[38] Thus, bias attributable to heterogeneity in the definition and period of exposure is likely to be minimal.

Second, we acknowledge that SNP rs2965667 and the highly correlated rs10505806 are relatively rare and imputed in all studies. However, we directly genotyped rs10505806 in cases and controls within 2 cohorts included in our study population. The high overall concordance (r = 0.89) between imputed and directly genotyped data and the consistent gene × environment interaction statistical effect using either imputed or directly genotyped data support our assumption that our results are not greatly affected by the amount of imputed data.

Although prior GWAS-based studies have traditionally examined promising findings within a replication cohort, we did not split our data into discovery and replication sets because the most powerful analytical approach is a combined analysis across all studies.[39] This approach is increasingly used as more individual-level GWAS data are becoming available.[40] Moreover, the consistency of our findings and lack of heterogeneity across distinct study cohorts supports the validity of the results.

Conclusions

In this genome-wide investigation of gene × environment interactions, use of aspirin, NSAIDs, or both was associated with lower risk of colorectal cancer, and the association of these medications with colorectal cancer risk differed according to genetic variation at 2 SNPs at chromosomes 12 and 15. Validation of these findings in additional populations may facilitate targeted colorectal cancer prevention strategies.

ARTICLE INFORMATION

Author Affiliations: Department of Epidemiology, Richard M. Fairbanks School of Public Health, Indiana University, Indianapolis (Nan); Indiana University Melvin and Bren Simon Cancer Center, Indianapolis (Nan); National Human Genome Research Institute, National Institutes of Health, Bethesda, Maryland (Hutter, Seminara); Public Health Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, Washington (Lin, Ulrich, White, Carlson, Gong, Harrison, Jiao, Newcomb, Pflugeisen, Potter, Qu, Rosse, Thornquist, Warnick, Peters, Hsu); Epidemiology Research Program, American Cancer Society, Atlanta, Georgia (Jacobs, Campbell); Huntsman Cancer Institute, University of Utah, Salt Lake City (Ulrich); Department of Epidemiology, University of Washington School of Public Health, Seattle (White, Peters); Division of Gastroenterology and Hepatology, University of North Carolina School of Medicine, Chapel Hill (Baron); Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, Bethesda, Maryland (Berndt, Chanock); Division of Clinical Epidemiology and Aging Research, German Cancer Research Center (DKFZ), Heidelberg, Germany (Brenner, Butterbach, Hoffmeister); German Cancer Consortium (DKTK), Heidelberg, Germany (Brenner); Division of Research, Kaiser Permanente Medical Care Program of Northern California, Oakland (Caan); Department of Preventive Medicine, Keck School of Medicine, University of Southern California, Los Angeles (Casey, Figueiredo, Haile, Schumacher, Gauderman); Division of Cancer Epidemiology, German Cancer Research Center, Heidelberg, Germany (Chang-Claude, Rudolph); Prevention and Cancer Control, Cancer Care Ontario, Toronto, Ontario, Canada (Cotterchio); Genetic Basis of Human Disease Division, Translational Genomics Research Institute (TGen), Phoenix, Arizona (Duggan); Department of Medical Oncology, Dana Farber Cancer Institute, Boston, Massachusetts (Fuchs, Ogino); Channing Division of Network Medicine, Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, Massacusetts (Giovannucci, Chan); Division of Epidemiology, Department of Population Health, New York University School of Medicine, New York, New York (Hayes); Melbourne School of Population Health, University of Melbourne, Victoria, Australia (Hopper, Jenkins); Department of Medical Biophysics, University of Toronto, Toronto, Ontario, Canada (Hudson); Ontario Institute for Cancer Research, Toronto, Ontario, Canada (Hudson, Lemire); Department of Health Sciences Research, Mayo Clinic, Scottsdale, Arizona (Lindor); Epidemiology Program, University of Hawaii Cancer Center, Honolulu (Le Marchand); Department of Pathology, Brigham and Women's Hospital, Boston, Massachusetts (Ogino); Department of Epidemiology, Harvard School of Public Health, Boston, Masschusetts (Ogino); Centre for Public Health Research, Massey University, Wellington, New Zealand (Potter); Department of Medicine and Epidemiology, University of Pittsburgh Medical Center, Pittsburgh, Pennsylvania (Schoen); Department of Internal Medicine, University of Utah Health Sciences Center, Salt Lake City (Slattery); Department of Laboratory Medicine and Pathology, Mayo Clinic, Rochester, Minnesota (Thibodeau); Department of Medical Genetics, Mayo Clinic, Rochester, Minnesota (Thibodeau); Division of Biostatistics and Epidemiology, Department of Preventive Medicine, University of Tennessee Health Science Center, Memphis (Thomas); Clinical Epidemiology Program, Ottawa Hospital Research Institute, Ottawa, Ontario, Canada (Zanke); Department of Biostatistics, University of Washington, Seattle (Hsu); Division of Gastroenterology, Massachusetts General Hospital and Harvard Medical School, Boston, Massachusetts (Chan).

Author Contributions: Ms Lin and Dr Peters had full access to all of the data in the study and take responsibility for the integrity of the data and the accuracy of the data analysis.
Drs Peters, Hsu, and Chan share senior authorship.
Study concept and design: Nan, Hutter, Ulrich, Brenner, Campbell, Chang-Claude, Fuchs, Hopper, Seminara, Slattery, Thibodeau, Peters, Hsu, Chan.
Acquisition, analysis, or interpretation of data: Nan, Hutter, Lin, Jacobs, Ulrich, White, Baron, Berndt, Brenner, Butterbach, Caan, Campbell, Carlson, Casey, Chanock, Cotterchio, Duggan, Figueiredo, Fuchs, Giovannucci, Gong, Haile, Harrison, Hayes, Hoffmeister, Hopper, Hudson, Jenkins, Jiao, Lindor, Lemire, Le Marchand, Newcomb, Ogino, Pflugeisen, Potter, Qu, Rosse, Rudolph, Schoen, Schumacher, Slattery, Thomas, Thornquist, Warnick, Zanke, Gauderman, Peters, Hsu, Chan.
Drafting of the manuscript: Nan, Lin, Campbell, Chang-Claude, Figueiredo, Fuchs, Warnick, Peters, Hsu, Chan.
Critical revision of the manuscript for important intellectual content: Nan, Hutter, Jacobs, Ulrich, White, Baron, Berndt, Brenner, Butterbach, Caan, Carlson, Casey, Chanock, Cotterchio, Duggan, Fuchs, Giovannucci, Gong, Haile, Harrison, Hayes, Hoffmeister, Hopper, Hudson, Jenkins, Jiao, Lindor, Lemire, Le Marchand, Newcomb, Ogino, Pflugeisen, Potter, Qu, Rosse, Rudolph, Schoen, Schumacher, Seminara, Slattery, Thibodeau, Thomas, Thornquist, Zanke, Gauderman, Peters, Hsu, Chan.
Statistical analysis: Nan, Hutter, Lin, Campbell, Chang-Claude, Fuchs, Jiao, Lemire, Pflugeisen, Qu, Gauderman, Hsu, Chan.
Obtained funding: Brenner, Caan, Chanock, Fuchs, Giovannucci, Haile, Hopper, Hudson, Jenkins, Lindor, Newcomb, Schoen, Slattery, Peters, Chan.
Administrative, technical, or material support: White, Brenner, Campbell, Duggan, Gong, Harrison, Hayes, Hoffmeister, Hopper, Newcomb, Potter, Rudolph, Schumacher, Thibodeau, Thornquist, Warnick, Zanke, Chan.
Study supervision: Nan, Brenner, Campbell, Chanock, Fuchs, Harrison, Hudson, Potter, Peters, Hsu, Chan.

Conflict of Interest Disclosures: All authors have completed and submitted the ICMJE Form for Disclosure of Potential Conflicts of Interest. Dr

Baron reported holding a use patent for aspirin as a colorectal chemopreventive agent. Dr Zanke reported holding a patent licensed to Arctic Dx. Dr Chan reported receiving personal fees from Bayer Healthcare, Pozen, and Pfizer. No other authors reported disclosures.

Funding/Support: GECCO (Genetics and Epidemiology of Colorectal Cancer Consortium) was supported by the National Cancer Institute (NCI), National Institutes of Health (NIH), US Department of Health and Human Services (U01 CA137088; R01 CA059045). CCFR (Colon Cancer Family Registry) was supported by the NIH (RFA CA-95-011) and through cooperative agreements with members of the Colon Cancer Family Registry and prinicipal investigators. This genome wide scan was supported by the NCI, NIH, by U01 CA122839. The following Colon CFR centers contributed data to this manuscript and were supported by NIH: Australasian Colorectal Cancer Family Registry (U01 CA097735), Ontario Registry for Studies of Familial Colorectal Cancer (U01 CA074783), and Seattle Colorectal Cancer Family Registry (U01 CA074794). DACHS (Darmkrebs: Chancen der Verhütung durch Screening Study) is supported by the German Research Council (Deutsche Forschungsgemeinschaft, BR 1704/6-1, BR 1704/6-3, BR 1704/6-4 and CH 117/1-1) and the German Federal Ministry of Education and Research (01KH0404 and 01ER0814). DALS (Diet, Activity and Lifestyle Study) is suported by the NIH (R01 CA48998 to Dr Slattery). HPFS (Health Professionals Follow-up Study) is supported by the NIH (P01 CA 055075, UM1 CA167552, R01 137178, R01 CA151993, and P50 CA 127003) and the NHS (Nurses' Health Study) by the NIH (R01 CA137178, P01 CA 087969, R01 CA151993, and P50 CA 127003). Dr Chan is also supported by K24 DK098311 and is a Damon Runyon Clinical Investigator. OFCCR (Ontario Familial Colorectal Cancer Registry) is supported by the NIH through funding allocated to the Ontario Registry for Studies of Familial Colorectal Cancer (U01 CA074783) (see CCFR section above). As subset of ARCTIC, OFCCR is supported by a GL2 grant from the Ontario Research Fund, the Canadian Institutes of Health Research, and the Cancer Risk Evaluation (CaRE) Program grant from the Canadian Cancer Society Research Institute. Dr Hudson and Dr Zanke are recipients of Senior Investigator Awards from the Ontario Institute for Cancer Research, through generous support from the Ontario Ministry of Research and Innovation. PLCO (Prostate, Lung, Colorectal and Ovarian Cancer Screening Trial) is supported by the Intramural Research Program of the Division of Cancer Epidemiology and Genetics and by contracts from the Division of Cancer Prevention, NCI, NIH, Department of Health and Human Services. PMH-CCFR (Postmenopausal Hormone Study–Colon Cancer Family Registry) is suported by the NIH (R01 CA076366 to Dr Newcomb). VITAL (VITamins And Lifestyle) is supported by the NIH (K05 CA154337). WHI (Women's Health Initiative) is supported by the National Heart, Lung, and Blood Institute, NIH, US Department of Health and Human Services, through contracts HHSN268201100046C, HHSN268201100001C, HHSN268201100002C, HHSN268201100003C, HHSN268201100004C, and HHSN271201100004C.

Role of the Funders/Sponsors: The funders and sponsors had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication.

Disclaimer: The content of this manuscript does not necessarily reflect the views or policies of the NCI or any of the collaborating centers in the CFRs, nor does mention of trade names, commercial products, or organizations imply endorsement by the US government or the CFR.

Additional Contributions: We thank all study participants for making this work possible. We appreciate the efforts of the GECCO Coordinating Center ensuring the success of this collaboration. For DACHS, we thank all participants and cooperating clinicians, and Ute Handte-Daub, Renate Hettler-Jensen, Utz Benscheid, Muhabbet Celik, and Ursula Eilber for excellent technical assistance. For NHS and HPFS, we acknowledge all participants and staff; Patrice Soule and Hardeep Ranu of the Dana Farber Harvard Cancer Center High-Throughput Polymorphism Core, who assisted in the genotyping under the supervision of Dr Immaculata Devivo and Dr David Hunter; Qun Guo, who assisted in programming. We also thank the cancer registries of Alabama, Arizona, Arkansas, California, Colorado, Connecticut, Delaware, Florida, Georgia, Idaho, Illinois, Indiana, Iowa, Kentucky, Louisiana, Maine, Maryland, Massachusetts, Michigan, Nebraska, New Hampshire, New Jersey, New York, North Carolina, North Dakota, Ohio, Oklahoma, Oregon, Pennsylvania, Rhode Island, South Carolina, Tennessee, Texas, Virginia, Washington, and Wyoming. For PLCO, we thank Drs Christine Berg and Philip Prorok, Division of Cancer Prevention, National Cancer Institute; the Screening Center investigators and staff for the PLCO Cancer Screening Trial; Tom Riley and staff, Information Management Services Inc; Barbara O'Brien and staff, Westat Inc; and Drs Bill Kopp, Wen Shao, and staff, SAIC-Frederick. For PMH, we thank the staff of the Hormones and Colon Cancer study. For WHI, we thank the WHI investigators and staff for their dedication, and the study participants for making the program possible. A full listing of WHI investigators can be found on the WHI website. Last, we acknowledge COMPASS (Comprehensive Center for the Advancement of Scientific Strategies) at the Fred Hutchinson Cancer Research Center for its work harmonizing the GECCO epidemiologic data set.

Correction: This article was corrected online on March 25, 2015, to add a note regarding shared authorship, correct an author degree, and correct a typographical error in the Acknowledgment section.

REFERENCES

1. Chan AT, Giovannucci EL, Meyerhardt JA, Schernhammer ES, Curhan GC, Fuchs CS. Long-term use of aspirin and nonsteroidal anti-inflammatory drugs and risk of colorectal cancer. *JAMA*. 2005;294(8):914-923.

2. Flossmann E, Rothwell PM; British Doctors Aspirin Trial and the UK-TIA Aspirin Trial. Effect of aspirin on long-term risk of colorectal cancer: consistent evidence from randomised and observational studies. *Lancet*. 2007;369(9573):1603-1613.

3. Friis S, Poulsen AH, Sørensen HT, et al. Aspirin and other non-steroidal anti-inflammatory drugs and risk of colorectal cancer: a Danish cohort study. *Cancer Causes Control*. 2009;20(5):731-740.

4. García Rodríguez LA, Huerta-Alvarez C. Reduced incidence of colorectal adenoma among long-term users of nonsteroidal antiinflammatory drugs: a pooled analysis of published studies and a new population-based study. *Epidemiology*. 2000;11(4):376-381.

5. Rothwell PM, Wilson M, Elwin CE, et al. Long-term effect of aspirin on colorectal cancer incidence and mortality: 20-year follow-up of five randomised trials. *Lancet*. 2010;376(9754):1741-1750.

6. Hutter CM, Chang-Claude J, Slattery ML, et al. Characterization of gene-environment interactions for colorectal cancer susceptibility loci. *Cancer Res*. 2012;72(8):2036-2044.

7. Hutter CM, Slattery ML, Duggan DJ, et al. Characterization of the association between 8q24 and colon cancer: gene-environment exploration and meta-analysis. *BMC Cancer*. 2010;10:670.

8. Makar KW, Poole EM, Resler AJ, et al. COX-1 (*PTGS1*) and COX-2 (*PTGS2*) polymorphisms, NSAID interactions, and risk of colon and rectal cancers in two independent populations. *Cancer Causes Control*. 2013;24(12):2059-2075.

9. Nan H, Morikawa T, Suuriniemi M, et al. Aspirin use, 8q24 single nucleotide polymorphism rs6983267, and colorectal cancer according to CTNNB1 alterations. *J Natl Cancer Inst*. 2013;105(24):1852-1861.

10. Seufert BL, Poole EM, Whitton J, et al. *IκBKβ* and *NFκB1*, NSAID use and risk of colorectal cancer in the Colon Cancer Family Registry. *Carcinogenesis*. 2013;34(1):79-85.

11. Jiao S, Hsu L, Hutter CM, Peters U. The use of imputed values in the meta-analysis of genome-wide association studies. *Genet Epidemiol*. 2011;35(7):597-605.

12. Cochran WG. The combination of estimates from different experiments. *Biometrics*. 1954;10:101-129.

13. Pfeiffer RM, Gail MH, Pee D. On combining data from genome-wide association studies to discover disease-associated SNPs. *Stat Sci*. 2009;24(4):547-560.

14. Evangelou E, Ioannidis JP. Meta-analysis methods for genome-wide association studies and beyond. *Nat Rev Genet*. 2013;14(6):379-389.

15. Begum F, Ghosh D, Tseng GC, Feingold E. Comprehensive literature review and statistical considerations for GWAS meta-analysis. *Nucleic Acids Res*. 2012;40(9):3777-3784.

16. Prage EB, Pawelzik SC, Busenlehner LS, et al. Location of inhibitor binding sites in the human inducible prostaglandin E synthase, MPGES1. *Biochemistry*. 2011;50(35):7684-7693.

17. Morgenstern R, Zhang J, Johansson K. Microsomal glutathione transferase 1: mechanism and functional roles. *Drug Metab Rev*. 2011;43(2):300-306.

18. Nakanishi M, Gokhale V, Meuillet EJ, Rosenberg DW. mPGES-1 as a target for cancer suppression: a comprehensive invited review "Phospholipase A2 and lipid mediators." *Biochimie*. 2010;92(6):660-664.

19. Castellone MD, Teramoto H, Gutkind JS. Cyclooxygenase-2 and colorectal cancer chemoprevention: the beta-catenin connection. *Cancer Res*. 2006;66(23):11085-11088.

20. Castellone MD, Teramoto H, Williams BO, Druey KM, Gutkind JS. Prostaglandin E2 promotes colon cancer cell growth through a Gs-axin-beta-catenin signaling axis. *Science*. 2005; 310(5753):1504-1510.

21. Kamei D, Murakami M, Nakatani Y, Ishikawa Y, Ishii T, Kudo I. Potential role of microsomal prostaglandin E synthase-1 in tumorigenesis. *J Biol Chem*. 2003;278(21):19396-19405.

22. Murakami M, Naraba H, Tanioka T, et al. Regulation of prostaglandin E2 biosynthesis by inducible membrane-associated prostaglandin E2 synthase that acts in concert with cyclooxygenase-2. *J Biol Chem*. 2000;275(42): 32783-32792.

23. Sherratt PJ, McLellan LI, Hayes JD. Positive and negative regulation of prostaglandin E2 biosynthesis in human colorectal carcinoma cells by cancer chemopreventive agents. *Biochem Pharmacol*. 2003;66(1):51-61.

24. Dihlmann S, Siermann A, von Knebel Doeberitz M. The nonsteroidal anti-inflammatory drugs aspirin and indomethacin attenuate beta-catenin/TCF-4 signaling. *Oncogene*. 2001;20 (5):645-653.

25. Gala MK, Chan AT. Molecular pathways: aspirin and *Wnt* signaling—a molecularly targeted approach to cancer prevention and treatment [published online December 11, 2014]. *Clin Cancer Res*. doi: 10.1158/1078-0432.CCR-14-0877.

26. Kwon YJ, Lee SJ, Koh JS, et al. Genome-wide analysis of DNA methylation and the gene expression change in lung cancer. *J Thorac Oncol*. 2012;7(1):20-33.

27. Aoyama M, Ozaki T, Inuzuka H, et al. LMO3 interacts with neuronal transcription factor, HEN2, and acts as an oncogene in neuroblastoma. *Cancer Res*. 2005;65(11):4587-4597.

28. Fresno Vara JA, Casado E, de Castro J, Cejas P, Belda-Iniesta C, González-Barón M. PI3K/Akt signalling pathway and cancer. *Cancer Treat Rev*. 2004;30(2):193-204.

29. Kaur J, Sanyal SN. PI3-kinase/Wnt association mediates COX-2/PGE(2) pathway to inhibit apoptosis in early stages of colon carcinogenesis: chemoprevention by diclofenac. *Tumour Biol*. 2010; 31(6):623-631.

30. Liao X, Lochhead P, Nishihara R, et al. Aspirin use, tumor *PIK3CA* mutation, and colorectal-cancer survival. *N Engl J Med*. 2012;367(17):1596-1606.

31. Domingo E, Church DN, Sieber O, et al. Evaluation of *PIK3CA* mutation as a predictor of benefit from nonsteroidal anti-inflammatory drug therapy in colorectal cancer. *J Clin Oncol*. 2013;31 (34):4297-4305.

32. Azimzadeh P, Romani S, Mohebbi SR, et al. Interleukin-16 (*IL-16*) gene polymorphisms in Iranian patients with colorectal cancer. *J Gastrointestin Liver Dis*. 2011;20(4):371-376.

33. Cruikshank WW, Kornfeld H, Center DM. Interleukin-16. *J Leukoc Biol*. 2000;67(6):757-766.

34. Gerhard R, Queisser S, Tatge H, et al. Down-regulation of interleukin-16 in human mast cells HMC-1 by *Clostridium difficile* toxins A and B. *Naunyn Schmiedebergs Arch Pharmacol*. 2011;383 (3):285-295.

35. Grivennikov SI, Karin M. Inflammatory cytokines in cancer: tumour necrosis factor and interleukin 6 take the stage. *Ann Rheum Dis*. 2011; 70(suppl 1):i104-i108.

36. Klampfer L. Cytokines, inflammation and colon cancer. *Curr Cancer Drug Targets*. 2011;11(4):451-464.

37. Symmons O, Uslu VV, Tsujimura T, et al. Functional and topological characteristics of mammalian regulatory domains. *Genome Res*. 2014;24(3):390-400.

38. Chan AT, Arber N, Burn J, et al. Aspirin in the chemoprevention of colorectal neoplasia: an overview. *Cancer Prev Res (Phila)*. 2012;5(2):164-178.

39. Skol AD, Scott LJ, Abecasis GR, Boehnke M. Joint analysis is more efficient than replication-based analysis for two-stage genome-wide association studies. *Nat Genet*. 2006;38(2):209-213.

40. Pearce CL, Rossing MA, Lee AW, et al; Australian Cancer Study (Ovarian Cancer); Australian Ovarian Cancer Study Group; Ovarian Cancer Association Consortium. Combined and interactive effects of environmental and GWAS-identified risk factors in ovarian cancer. *Cancer Epidemiol Biomarkers Prev*. 2013;22(5):880-890.