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.

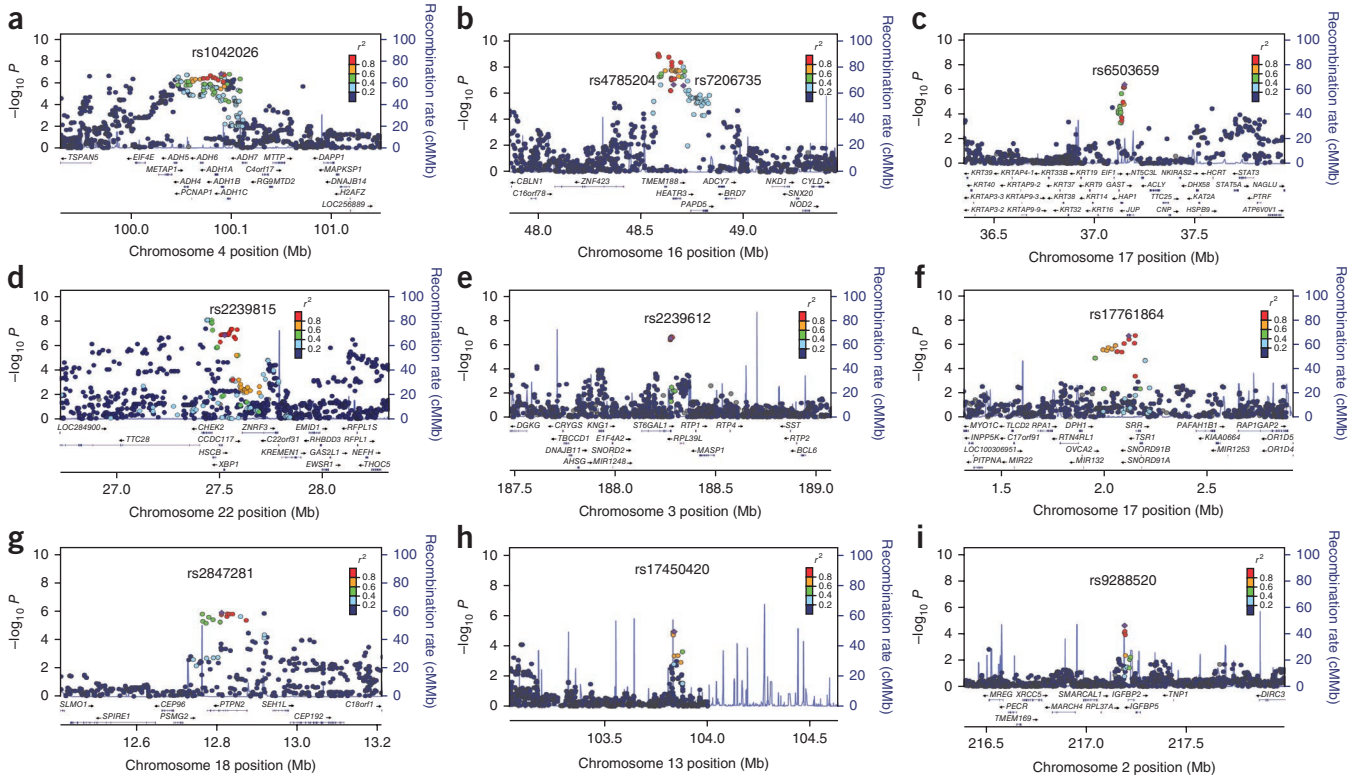

Figure 1 Regional plots of the association results for genotyped and imputed SNPs and the recombination rates within nine significant susceptibility loci. (**a**–**i**) The significant loci are located in chromosomes 4q23 (**a**),16q12.1 (**b**),17q21 (**c**), 22q12 (**d**), 3q27 (**e**), 17p13 (**f**), 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

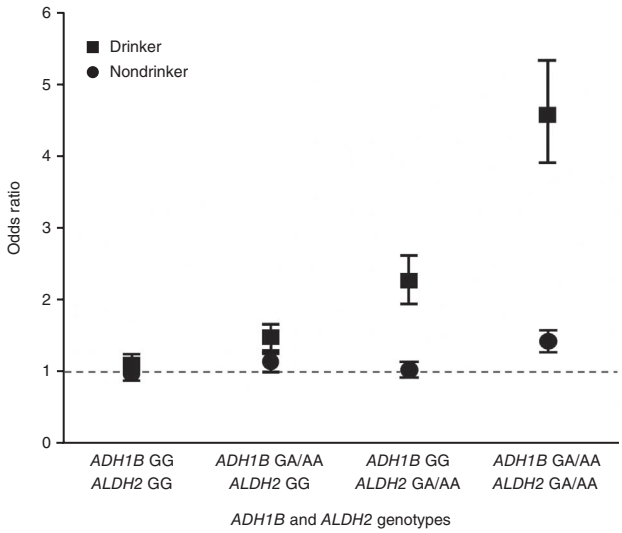

Figure 2 Plots showing the ORs for ESCC in alcohol drinkers and nondrinkers with different *ADH1B* rs1042026 and *ALDH2* rs11066015 genotypes. The vertical bars represent the 95% CIs. The horizontal dashed line indicates the null value (OR = 1.0).

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 nonreceptor 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).