



ORIGINAL ARTICLE

Genetic susceptibility to hepatocellular carcinoma in chromosome 22q13.31, findings of a genome-wide association study

Zhanwei Wang,* Anuradha S Budhu,[†] Yi Shen,* Linda Lou Wong,*  Brenda Y Hernandez,* Maarit Tiirikainen,* Xiaomei Ma,[‡] Melinda L Irwin,[‡] Lingeng Lu,[‡] Hongyu Zhao,[‡] Joseph K Lim,[§] Tamar Taddei,[§] Lopa Mishra,[¶] Karen Pawlish,[¶] Antoinette Stroup,** Robert Brown,^{††} Mindie H Nguyen,^{‡‡} Jill Koshiol,^{§§} Maria O Hernandez,^{¶¶} Marshonna Forgues,^{¶¶} Hwai-I Yang,^{¶¶¶} Mei-Hsuan Lee,^{¶¶¶} Yu-Han Huang,^{¶¶¶} Motoki Iwasaki,^{¶§} Atsushi Goto,^{¶§} Shiori Suzuki,^{¶§} Koichi Matsuda,^{¶¶} Chizu Tanikawa,^{¶¶} Yoichiro Kamatani,^{¶¶} Dean Mann,^{¶¶} Maria Guarnera,^{¶¶} Kirti Shetty,^{¶¶¶} Claire E Thomas,^{¶¶¶} Jian-Min Yuan,^{¶¶¶} Chiea Chuen Khor,^{†††} Woon-Puay Koh,^{†††} Harvey Risch,[‡] Xin Wei Wang[†] and Herbert Yu* 

*University of Hawaii Cancer Center, Honolulu, Hawaii, [†]Laboratory of Human Carcinogenesis, Liver Cancer Program, Center for Cancer Research ^{§§}Division of Cancer Epidemiology and Genetics, National Cancer Institute, ^{¶¶}Laboratory of Human Carcinogenesis, Center for Cancer Research, National Cancer Institute, Bethesda, Departments of ^{¶¶}Pathology, ^{¶¶¶}Gastroenterology and Hepatology, University of Maryland School of Medicine, Baltimore, Maryland, [‡]Yale School of Public Health, [§]Yale School of Medicine, New Haven, Connecticut, [¶]Center for Translational Medicine, Department of Surgery, The George Washington University, Washington, District of Columbia, ^{¶¶}New Jersey State Cancer Registry, New Jersey Department of Health, Trenton, ^{¶¶}Rutgers Cancer Institute, and Rutgers School of Public Health, New Brunswick, New Jersey, ^{‡‡}Division of Gastroenterology and Hepatology, Stanford University Medical Center, Palo Alto, California, ^{¶¶¶}Division of Cancer Control and Population Sciences, University of Pittsburgh Medical Center (UPMC) Hillman Cancer Center, ^{¶¶¶}Department of Epidemiology, Graduate School of Public Health, University of Pittsburgh, Pittsburgh, Pennsylvania, USA, ^{¶¶¶}Genomics Research Center, Academia Sinica, ^{¶¶¶}Institute of Clinical Medicine, National Yang Ming University, Taipei, Taiwan, ^{¶§}Division of Epidemiology, Center for Public Health Sciences, National Cancer Center, ^{¶¶}Graduate School of Frontier Sciences, and Institute of Medical Science, University of Tokyo, Tokyo, Japan, ^{††}Genome Institute of Singapore, Agency for Science, Technology and Research, ^{¶¶}Singapore Eye Research Institute, ^{††}Health Systems and Services Research, Duke-NUS Medical School Singapore, ^{‡‡}Saw Swee Hock School of Public Health, National University of Singapore, Singapore, Singapore

Key words

genome-wide association study, liver cancer, non-alcoholic fatty liver disease, *PNPLA3*, *SAMM50*.

Accepted for publication 7 November 2021.

Correspondence

Herbert Yu, University of Hawaii Cancer Center, 701 Ilalo Street, Honolulu, HI 96813, USA.
Email: hyu@cc.hawaii.edu
Xin Wei Wang, National Cancer Institute, 37 Convent Drive, Bethesda, MD 20892, USA.
Email: xw3u@nih.gov

Declaration of conflict of interest: The authors declare no conflicts of interest that pertain to this work.

Author contribution: Zhanwei Wang, Anuradha S Budhu, and Yi Shen contributed equally to the work.

Financial support: The study was supported by NIH grants R01CA138698 to Herbert Yu, U01230690 to Lopa Mishra, R01CA144034, and UM1CA182876 to Jian-Min Yuan. The Genomics Shared Resource at University of Hawaii Cancer Center (RRID:SCR 019085) is supported by NIH grant P30CA071789. Anuradha S Budhu, Maria O Hernandez, Marshonna Forgues, and Xin Wei

Abstract

Background and Aim: Chronic hepatitis C virus (HCV) infection, long-term alcohol use, cigarette smoking, and obesity are the major risk factors for hepatocellular carcinoma (HCC) in the United States, but the disease risk varies substantially among individuals with these factors, suggesting host susceptibility to and gene–environment interactions in HCC. To address genetic susceptibility to HCC, we conducted a genome-wide association study (GWAS).

Methods: Two case-control studies on HCC were conducted in the United States. DNA samples were genotyped using the Illumina microarray chip with over 710 000 single nucleotide polymorphisms (SNPs). We compared these SNPs between 705 HCC cases and 1455 population controls for their associations with HCC and verified our findings in additional studies.

Results: In this GWAS, we found that two SNPs were associated with HCC at $P < 5E-8$ and six SNPs at $P < 5E-6$ after adjusting for age, sex, and the top three principal components (PCs). Five of the SNPs in chromosome 22q13.31, three in *PNPLA3* (rs2281135, rs2896019, and rs4823173) and two in *SAMM50* (rs3761472, rs3827385), were replicated in a small US case-control study and a cohort study in Singapore. The associations remained significant after adjusting for body mass index and HCV infection. Meta-analysis of multiple datasets indicated that these SNPs were significantly associated with HCC.

Conclusions: SNPs in *PNPLA3* and *SAMM50* are known risk loci for non-alcoholic fatty liver disease (NAFLD) and are suspected to be associated with HCC. Our GWAS demonstrated the associations of these SNPs with HCC in a US population. Biological mechanisms underlying the relationship remain to be elucidated.

Wang were supported by grants ZIA-BC10313, ZIA-BC010876, and ZIA-BC010877 from the Intramural Research Program of the Center for Cancer Research, National Cancer Institute. The work in Singapore was supported by a grant from the Singapore National Medical Research Council (NMRC/CIRG/1456/2016). The Japan Public Health Center-based Prospective Study was supported by the National Cancer Center Research and Development Fund since 2011 and a grant-in-aid for Cancer Research from the Ministers of Health, Labor, and Welfare of Japan from 1989 to 2010.

Funding support: National Institutes of Health P30CA071789 R01CA138698 U01230690

Introduction

Liver cancer is one of the most common malignancies and a leading cause of cancer death worldwide.^{1,2} In the United States, hepatocellular carcinoma (HCC) incidence has been rising steadily over the past three decades, from 2.7/100 000 in 1975–1980 to 8.3/100 000 in 2012–2016.³ HCC is the major histological type of liver cancer. The main risk factors for HCC in the United States include chronic hepatitis C virus (HCV) infection, heavy alcohol drinking, cigarette smoking, obesity (or metabolic disorders), diabetes, nonalcoholic fatty liver disease (NAFLD), and certain genetic disorders (e.g. hereditary hemochromatosis).⁴ Using the SEER data, Makarova-Rusher *et al.* estimated that these risk factors accounted for 60% population attributable risk (PAR %) in the US elderly population. Moreover, the PAR% appeared to increase from 52% in 2000–2003 to 64% in 2008–2011, largely due to the rise of metabolic disorders.⁵ Even with the 64% of overall PAR%, more than a third of the HCC risk is still unaccounted which led to the speculation that genetic susceptibility may play a role in the development of HCC.

Several genome-wide association studies (GWAS) have found a number of single nucleotide polymorphisms (SNPs) in multiple genomic regions associated with HCC risk, including 1p36.22, 2q32.3, 6p21.32, 6p21.33, 6q15, 7q21.13, 8p12, 21q21.3, and 22q12.2.^{6–14} These studies were largely conducted among Asian populations where liver cancer is prevalent, and the major underlying causes are different, such as hepatitis B virus (HBV) infection and dietary exposure to aflatoxin B₁.¹⁵ So far, few GWAS have been reported in the United States. To address genetic susceptibility to HCC, we conducted a GWAS in the United States, and our findings were further evaluated in independent studies in the United States, Japan, Singapore, and Taiwan.

Materials and methods

Studies for genome-wide SNP analysis (phase 1).

HCC cases, patients with chronic liver diseases (CLD), and population controls were from two case-control studies, the Yale Liver Health Study (YLHS) and the National Cancer Institute/University of Maryland School of Medicine (NCI-UMD) study. YLHS was conducted between 2011 and 2016 and enrolled

731 cases and 1166 controls, of whom 501 cases and 989 controls provided saliva samples. The cases were identified through cancer registries and liver disease clinics in Connecticut ($n = 145$), New Jersey ($n = 236$), New York City ($n = 68$), Stanford/Palo Alto, California ($n = 46$), and Hawaii ($n = 6$). The controls were individuals without cancer randomly selected from the general populations in Connecticut ($n = 338$) and New Jersey ($n = 651$). Study details have been reported previously.¹⁶

The NCI-UMD study was registered with the government database (<https://clinicaltrials.gov/ct2/show/NCT00913757>). Study participants were recruited from the city of Baltimore, Maryland, between 2003 and 2016. Included in GWAS were 239 patients with HCC, 509 individuals with CLD due to hepatitis virus infection (86%), heavy alcohol drinking, NAFLD, or nonalcoholic steatohepatitis (NASH), and 512 population controls who lived in the neighborhoods of HCC cases without a history of liver disease. Both studies were approved by IRBs at all participating institutions and organizations. Each participant signed an informed consent. Personal information including demographic features and medical history was collected for the study.

Replication studies for selected SNPs (phase 2).

Two types of replications were conducted for validation. One was direct genotyping on the samples from a case-control study of HCC among non-Asians in Los Angeles County, California (the LA study). Details of the LA study were described previously.¹⁷

The second type of replication was conducted by analyzing the GWAS data generated in Japan, Singapore, and Taiwan. Seven GWAS datasets were analyzed including the Japan Public Health Center-based Prospective Study (JPHC) from the National Cancer Center of Japan,^{18,19} the BioBank Japan (BBJ) from the University of Tokyo^{7,20} which included two sets of comparison (HCC/HCV cases vs HCV controls [BBJ-HCV] and HCC/HBV cases vs HBV controls [BBJ-HBV]), the Singapore Chinese Health Study (Singapore) from the National University of Singapore,²¹ and the Taiwan Liver Cancer Network and Risk Evaluation of Viral Load Elevation and Associated Liver Disease/Cancer-Hepatitis B and C Virus (TLCN/REVEAL) from the National Yang-Ming University^{14,22} which included two sets of comparison (HCC/HCV cases vs HCV controls [Taiwan-HCV]

and HCC/HBV cases vs HBV controls [Taiwan-HBV]). Information on genotyping methods and numbers of cases and controls in each replication study is provided in Table S1, Supporting information.

DNA samples. YLHS collected saliva samples from study participants, using a saliva self-collection kit, OG-250 (DNA Genotek). Blood samples were collected from the participants in the NCI-UMD study and the LA study, and genomic DNA was extracted from buffy coats. Commercial kits were used for DNA extraction in both studies.

SNP genotyping. Infinium OmniExpress BeadChip (Illumina, San Diego, CA) was used for genotyping DNA samples from YLHS and NCI-UMD. The microarray chip contains more than 710 000 SNPs. After quality assessment, 2232 DNA samples including 721 HCC patients (482 from YLHS and 239 from NCI-UMD), 486 CLD controls (from NCI-UMD), and 1511 population controls (1008 from YLHS and 503 from NCI-UMD) were analyzed. The microarray assay was performed on 200 ng of genomic DNA, using the Illumina Infinium HTS Global Screening Arrays scanned on an Illumina iScan system at the Genomics Shared Resource at the University of Hawaii Cancer Center. The initial genotyping results were processed using the Illumina GenomeStudio software 2.0. Quality control procedures were performed using PLINK 2.0 (<http://www.cog-genomics.org/plink/2.0>).²³ We randomly selected 215 DNA samples (92 from YLHS and 123 from NCI-UMD) for blind repeat.

Table 1 Characteristics of HCC cases, chronic liver diseases (CLD) patients, and population controls in GWAS

Variables	Case (<i>n</i> = 705)	Control (<i>n</i> = 1455)	CLD* (<i>n</i> = 509)
Average age at diagnosis/ interview	62.8 ± 10.8	61.3 ± 12.2	58.6 ± 6.2
Gender			
Female	129 (18.3)	418 (28.7)	96 (18.9)
Male	576 (81.7)	1037 (71.3)	413 (81.1)
Race			
White	436 (62.1)	1065 (73.2)	173 (34.0)
Black	154 (21.9)	357 (24.5)	335 (65.8)
Others	112 (16.0)	33 (2.3)	1 (0.2)
Additional variables			
BMI	Case (<i>n</i> = 667)	Control (<i>n</i> = 1434)	CLD* (<i>n</i> = 508)
Normal (< 25)	240 (36.0)	408 (28.45)	160 (31.5)
Overweight (25–29)	256 (38.4)	579 (40.38)	177 (34.8)
Obese (≥ 30)	171 (25.6)	447 (31.17)	171 (33.7)
Viral status	Case (<i>n</i> = 579)	Control (<i>n</i> = 962)	CLD* (<i>n</i> = 324)
None	245 (42.6)	943 (98.0)	37 (11.4)
HCV positive	283 (48.6)	12 (1.3)	265 (81.8)
HCV/HBV positive	34 (5.8)	1 (0.1)	14 (4.3)
HBV positive	17 (2.9)	6 (0.6)	8 (2.5)

The results were highly concordant (99%). SNPs were excluded if they were: (i) on chromosome X, (ii) < 95% of genotyping call rate, (iii) < 0.05 of minor allele frequency (MAF), or (iv) not in Hardy–Weinberg equilibrium (HWE) ($P < 10^{-4}$).

Genetic relatedness of the DNA samples was examined using the pairwise identity-by-descent (IBD) analysis in PLINK 2.0, and population outlier and structure were evaluated with the Genome-wide Complex Trait Analysis (GCTA) (<http://cns.genomics.com/software/gcta/>).²⁴ In addition, DNA samples were excluded if their genotyping call rates were < 95%. After quality assessment, the final genotype data qualified for data analysis included 705 HCC cases, 486 CLD controls, and 1455

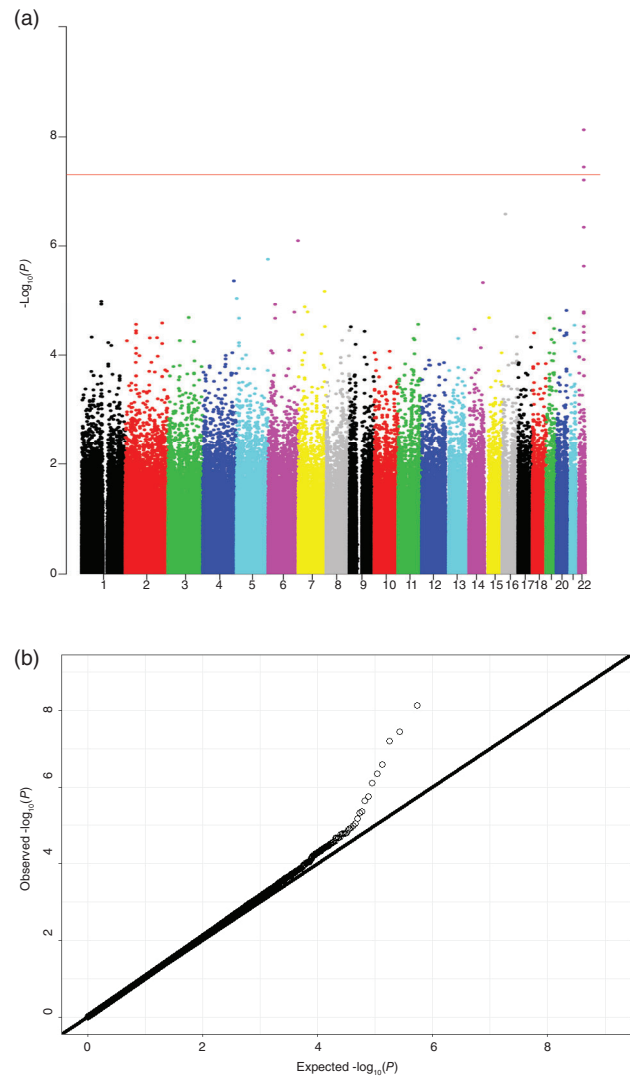


Figure 1 (a) Scatter plot of P values by 22 autosomal chromosomes in the GWAS (Manhattan plot). The P values were adjusted for age, gender, and the first three principle components. (b) Q-Q plot of the expected P values against the observed ones in the GWAS. The P values were adjusted for age, gender, and the first three principle components.

Table 2 Associations between HCC and top SNPs in GWAS using population controls

All samples									
SNP_ID	CHR	BP	REF	ALT	TEST	OR [†]	95% CI		P value
rs2281135	22	44332570	G	A	ADD	1.64	1.39	1.95	7.43E-09
rs2896019	22	44333694	A	C	ADD	1.60	1.35	1.89	3.63E-08
rs4823173	22	44328730	G	A	ADD	1.61	1.36	1.90	6.28E-08
rs10400971	16	17529169	G	A	ADD	1.90	1.49	2.43	2.64E-07
rs3761472	22	44368122	A	G	ADD	1.55	1.30	1.83	4.60E-07
rs9455680	6	169048842	G	A	ADD	1.64	1.35	2.00	8.01E-07
rs11750821	5	178634683	G	A	ADD	1.61	1.32	1.95	1.75E-06
rs3827385	22	44388817	A	G	ADD	1.49	1.26	1.76	2.34E-06
Samples with additional data for adjustment									
SNP_ID	CHR	BP	REF	ALT	TEST	OR [‡]	95% CI		P value
rs2281135	22	44332570	G	A	ADD	1.91	1.49	2.45	2.98E-07
rs2896019	22	44333694	A	C	ADD	1.89	1.47	2.42	5.13E-07
rs4823173	22	44328730	G	A	ADD	1.86	1.45	2.38	1.17E-06
rs10400971	16	17529169	G	A	ADD	0.66	0.39	1.12	0.1257
rs3761472	22	44368122	A	G	ADD	1.76	1.36	2.27	1.50E-05
rs9455680	6	169048842	G	A	ADD	1.64	1.21	2.24	0.0016
rs11750821	5	178634683	G	A	ADD	1.24	0.89	1.73	0.2021
rs3827385	22	44388817	A	G	ADD	1.51	1.18	1.92	1.00E-03

[†]Adjusted for age, gender, and 3 PCs.

[‡]Adjusted for age, gender, BMI, viral status, study site, and 3 PCs.

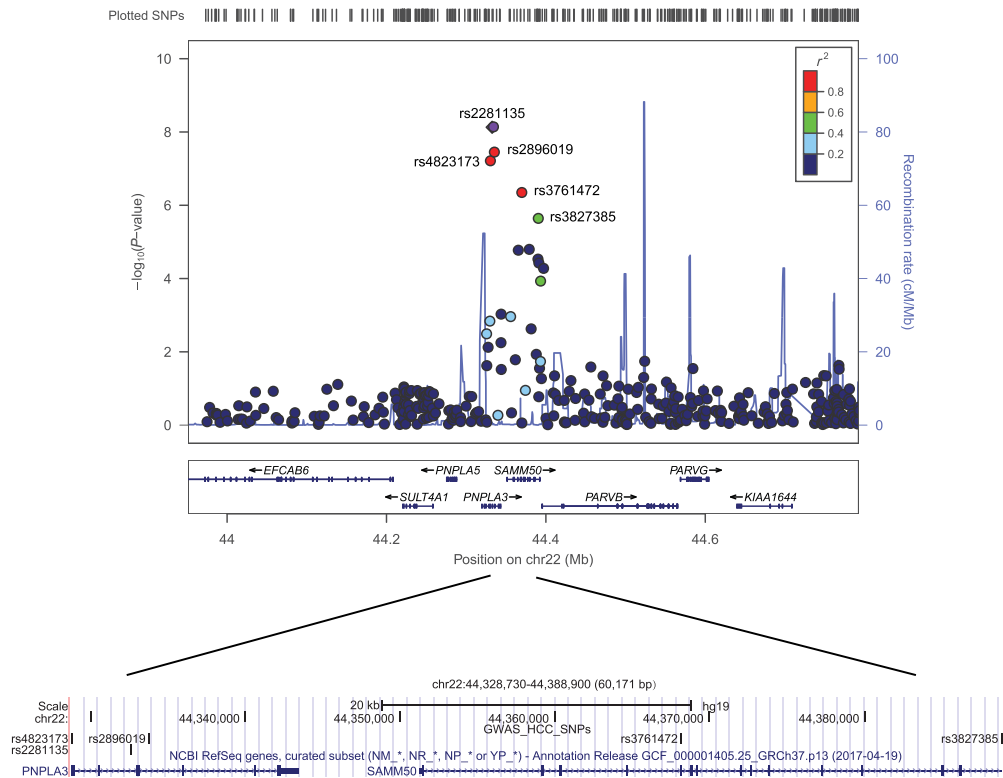


Figure 2 Scatter plot of the P values and locations of the SNPs in the region of chromosome 22q13.31 and their linkage disequilibrium (LD) with rs2281135. The P values were adjusted for age, gender, and the first three principle components.

population controls. R was used to plot the top three principal components (PCs) data (Fig. S1).

The SNP genotypes for replication in the LA study were determined using the TaqMan assay in the LightCycler 480 II real-time PCR system (Roche). Genomic DNA samples from 100 cases and 214 controls were available for replication.

Statistical analysis. Associations between SNPs and HCC risk were analyzed using an additive logistic regression model with initial adjustment for age, gender, and the top three PCs using PLINK 2.0. Additional adjustment was included subsequently on samples with information on BMI and HBV/HCV status. Subgroup analyses were also performed after the cases and controls were stratified by the status of HCV/HBV infection. Odds ratios (ORs) and 95% confidence intervals (CI) were calculated for the SNP association with HCC in the unconditional logistic regression using the homozygous wild type allele as reference. The Manhattan plot of $-\log_{10} P$ values and quantile-quantile plot were generated using the R software. LocusZoom was used to plot the top HCC-associated SNPs in a genomic region with linkage disequilibrium (LD) values and recombination rates calculated from 1000 Genomes (<http://locuszoom.org/>).²⁵

The Review Manager software (Revman version 5.3) was used for meta-analysis. Pooled OR and 95% CI were estimated using the random-effects model (the DerSimonian and Laird method).²⁶ We also used GTEx V8 (<https://www.gtexportal.org/home/>) to ascertain information on expression quantitative trait loci

(eQTL) which was used to estimate the SNP's influence on gene expression. RegulomeDB (<https://regulomedb.org/regulome-search/>) was used to predict the effect of a SNP on the gene function.²⁷ RegulomeDB annotates SNPs regarding their potential impacts on DNase, transcription factor binding, and promoter activity regulation.

Results

Phase 1 genome-wide analysis. Table 1 shows the demographic and clinical characteristics of study participants in phase 1 analysis. Compared to HCC cases, population controls were slightly younger and had greater percentages of females and Caucasians. A large majority of CLD patients (81.8%) were infected with HCV, and over 65% of CLD were African Americans. HBV infection was quite low in HCC cases (2.9%) and CLD patients (2.5%). Few population controls reported HCV or HBV infection, 1.3% and 0.6%, respectively. BMI did not seem to be substantially different among the groups.

After removing the disqualified SNPs, we had 536 522 polymorphisms in analysis. Figure 1a shows the P values for SNPs' association with HCC in the Manhattan plot. Two SNPs were found to be associated with HCC at a genome-wide level of significance ($< 5E-8$), and six had P values $< 5E-6$ after adjusting for age, gender and the top three PCs ($\lambda = 1.05$) (Table 2, upper panel: all samples). λ is the genetic inflation factor. These associations did not change substantially when we analyzed the data in

Table 3 Associations between HCC and top SNPs in GWAS using CLD[†] controls

SNP_ID	CHR	BP	REF	ALT	TEST	OR [‡]	95% CI		P value
rs2281135	22	44332570	G	A	ADD	1.32	1.07	1.57	0.026
rs2896019	22	44333694	A	C	ADD	1.27	1.00	1.62	0.054
rs4823173	22	44328730	G	A	ADD	1.30	1.05	1.55	0.044
rs10400971	16	17529169	G	A	ADD	NA [§]			
rs3761472	22	44368122	A	G	ADD	1.25	0.99	1.49	0.067
rs9455680	6	169048842	G	A	ADD	1.41	1.13	1.69	0.015
rs11750821	5	178634683	G	A	ADD	NA [§]			
rs3827385	22	44388817	A	G	ADD	1.33	1.09	1.57	0.022

[†]Chronic liver disease.

[‡]Adjusted for age, gender, and 3 PCs.

[§]Not available for analysis because the SNP was no longer qualified for analysis when CLD was used as controls.

Table 4 Associations between HCC and top SNPs in replication

SNP_ID	CHR	LA study		BBJ-HCV		BBJ-HBV		JPHC		Singapore		Taiwan-HBV		Taiwan-HCV	
		OR	P	OR	P	OR	P	OR	P	OR	P	OR	P	OR	P
rs2281135	22	2.02	0.002	1.09	0.099	1.00	0.979	1.13	0.31	1.26	0.025	1.09	0.17	1.14	0.11
rs2896019	22	1.87	0.003	1.09	0.099	1	0.999	1.14	0.28	1.26	0.027	NG		NG	
rs4823173	22	1.87	0.004	1.10	0.096	0.99	0.906	1.13	0.32	1.28	0.018	NG		NG	
rs10400971	16	1.65	0.230	1.29	0.071	1.16	0.632	1.14	0.65	0.79	0.640	NG		NG	
rs3761472	22	2.38	0.0004	1.04	0.486	1.06	0.620	1.14	0.27	1.24	0.045	NG		NG	
rs9455680	6	1.62	0.076	1.09	0.216	0.96	0.774	0.80	0.16	NG		NG		NG	
rs11750821	5	NG		0.81	0.147	0.68	0.244	0.47	0.08	NG		NG		NG	
rs3827385	22	2.02	0.003	1.03	0.595	1.07	0.552	1.18	0.17	1.22	0.061	NG		NG	

NG, not genotyped.

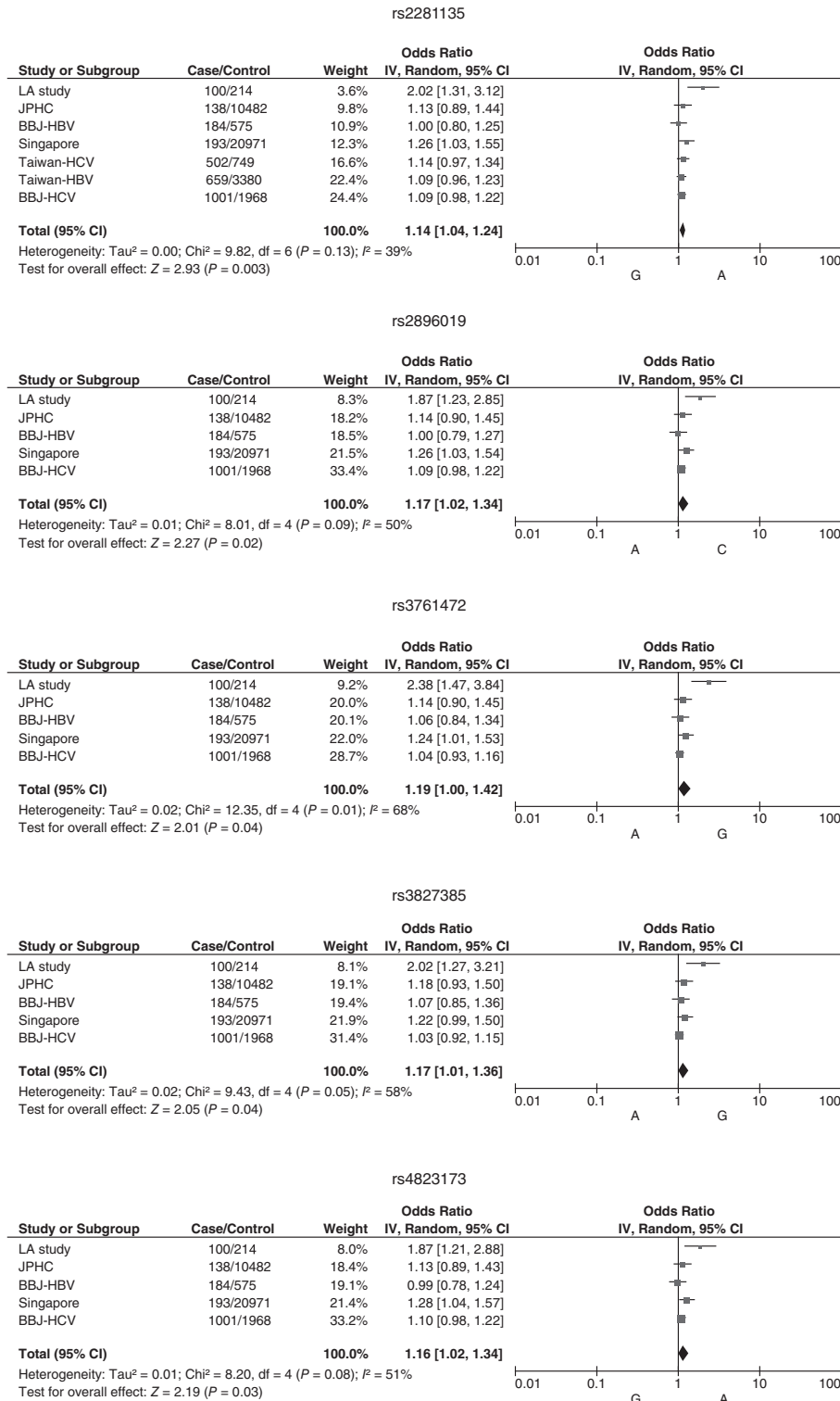


Figure 3 Results of meta-analysis on the replication studies with genotype data on the five SNPs in 22q13.31. Seven studies had genotype data available for rs2281135, and five studies had data for rs2896019, rs3761472, rs3827385, and rs4823173.

whites only, the largest racial group in phase 1 (data not shown). Figure 1b shows the Q-Q plot for all the observed and expected P values. Of the eight top SNPs, five were in chromosome

22 (rs2281135, rs2896019, rs4823173, rs3761472, and rs3827385), and one each in chromosomes 5 (rs11750821), 6 (rs9455680), and 16 (rs10400971). For the five SNPs in

chromosome 22, they were all located in the 22q13.31 regions (Fig. 2). Three SNPs (rs2281135, rs2896019, and rs4823173) were in high LD (pairwise $r^2 > 0.8$), while the other two (rs3761472 and rs3827385) were in a close link (pairwise $r^2 > 0.8$) (Table S2).

To control for risk factors, we performed additional analysis on samples with information on BMI and hepatitis virus infection. The associations of the five SNPs in 22q13.31 did not change substantially (Table 2, lower panel: samples with additional data for adjustment). Since many HCC developed in the background of CLD, we also compared HCC with CLD controls for the SNP association. The results showed that most of the SNPs in 22q13.31 were still significantly associated with HCC (Table 3), indicating that these SNPs were more relevant to HCC than CLD or had additional risk for HCC in the US population. We further analyzed these associations stratified by the status of hepatitis virus infection. The associations of HCC with SNPs in 22q13.31 remained significant in the study participants who were negative for viral infection but attenuated in those who were viral infection positive. This attenuation may be due to the small number of population controls with reported hepatitis virus infection. Results of these subgroup analyses are provided in Table S3.

Phase 2 replication and meta-analysis. Table 4 shows the results of our replication. Associations of HCC with the five SNPs in chromosome 22 found in our genome-wide analysis were all statistically significant in the LA study. Four of the 5 SNPs were also significant, and one was borderline significant in the Singapore cohort. Similar directions of associations were observed in most of the Japanese and Taiwanese studies although they were not statistically significant. HCC associations with SNPs in chromosomes 5, 6, and 16 were not shown in our replication studies. Meta-analysis of the SNPs in 22q13.31 was performed in the replication studies available (Fig. 3). For SNP rs2281135, we had genotype information from seven datasets. The summary OR for rs2281135 was 1.14 (95% CI: 1.04–1.24). For the rest of SNPs in 22q13.31, data were available from five datasets. The summary ORs for these SNPs were also statistically significant.

SNP eQTL and predicted function. Based on the genomic alignment (Fig. 2), SNPs rs2281135, rs4823173, and rs2896019 are in a gene named *PNPLA3* (Patatin-like phospholipase domain-containing protein 3), and SNPs rs3761472 and rs3827385 are in a nearby gene called *SAMM50* (Sorting and assembly machinery component 50 homolog). Both genes are known to be expressed in the liver (Fig. S2). RegulomeDB suggests that SNP rs3761472 is nonsynonymous, SNP rs3827385 resides in a transcription factor binding site, and four of the five SNPs, except rs2896019, may influence transcription factor binding (Table S4). Moreover, data from GTEx (version 8) show that the five SNPs in 22q13.31 are cis-eQTL for *SAMM50* in adipocytes of subcutaneous and visceral tissues, and the normalized magnitudes of association ranged between 0.13 and 0.23 (Table S5). These SNPs are also cis-eQTL for *PNPLA3* in the skin with normalized magnitudes ranging from -0.15 to -0.23 .

Discussion

In this study, we found five SNPs (rs2281135, rs2896019, rs4823173, rs3761472, and rs3827385) in 22q13.31 associated with HCC, and these associations did not change substantially after adjusting for BMI and hepatitis virus infection. The SNPs are highly linked within two nearby genomic regions where two coding genes reside, *PNPLA3* and *SAMM50*. To the best of our knowledge, no GWAS results on HCC have been reported for these SNPs in the US population. A recent study from France reported a different SNP in *PNPLA3*, rs738409, associated with HCC in people with CLD,²⁸ and this association has been found previously by a number of studies.^{29,30} SNP rs738409 is not on the microarray chip we used for our study. However, this SNP is in modest LD with the three *PNPLA3* SNPs in our study, r^2 ranging from 0.86 to 0.76. In a GWAS of NAFLD, Kawaguchi *et al.* found that SNP rs2896019 was associated with HCC among those with NASH. Although the number of NASH-HCC cases in that study was relatively small ($n = 58$), the finding did indicate an association between rs2896019 and HCC in a Japanese population.³¹ Their study also confirmed that the SNP was associated with NAFLD. Several previous GWAS have not found these SNPs associated with HCC. One of the possible explanations for this could be the study population. Most studies were done in Asia, like China, Japan, and Taiwan, where obesity is less prevalent and hepatitis virus infection is high compared to the United States and other western countries. Given that these SNPs are linked to NAFLD, a major outcome of obesity, it is quite possible that their associations with HCC are more readily detectable in areas where NAFLD and obesity are common.

Multiple studies have shown that SNPs in *PNPLA3* are associated with NAFLD. Romeo *et al.* were the first to report from a scan of nonsynonymous SNPs in a US population that SNP rs738409 in *PNPLA3* was associated with hepatic fat content, and the association was not affected by race/ethnicity, BMI, diabetes, and alcohol consumption.³² As a nonsynonymous SNP, rs738409 involves a nucleotide transversion which leads to an amino acid change from isoleucine to methionine (I148M). The observation by Romeo *et al.* was later confirmed by several GWAS reports. A GWAS by DiStefano *et al.* which included over 1800 individuals who provided data on histopathologic exam of liver biopsy showed that three SNPs in *PNPLA3* (rs4823173, rs2896019, and rs2281135) were associated with hepatic fat levels.³³ That study used a microarray chip same to ours, and the three SNPs associated with fatty liver were similar to our findings on HCC. A GWAS in Japan also found multiple SNPs in *PNPLA3* and *SAMM50*, including rs2896019 and rs3761472, associated with the risk of NAFLD.³⁴ The authors speculated that the entire block of *PNPLA3-SAMM50-PARVB* in 22q13.31 could be involved in the development and progression of NAFLD. A recent GWAS in a Korean population confirmed that NAFLD was associated with the known SNPs in *PNPLA3* and *SAMM50*, including rs2281135 and rs3761472.³⁵ A GWAS on serum levels of liver enzymes among the European populations showed that SNP rs2281135 in *PNPLA3* and SNP rs3761472 in *SAMM50* were associated with alanine aminotransferase (ALT) levels, suggesting a possible involvement of these genetic polymorphisms in liver diseases.³⁶ These observations were replicated by the Japanese GWAS mentioned above³⁴ as well as studies in Mexican Americans³⁷ and Chinese.³⁸

PNPLA3 encodes patatin-like phospholipase domain-containing protein 3 which is a membrane protein. *PNPLA3* was originally thought to be a triacylglycerol lipase, like *PNPLA2*, catalyzing the hydrolysis of triacylglycerol.³⁹ Unlike *PNPLA2* which is regulated by fasting, the activity of *PNPLA3* is regulated by feeding. Calorie intake and insulin increase its expression.^{40–42} In addition, different genotypes of *PNPLA3* were found to affect the function of *PNPLA3* protein in the liver.⁴³ Yang *et al.* showed that *PNPLA3* translated from the NAFLD risk alleles had a lower lipolytic activity than that from the non-risk alleles, confirming the functional relevance of SNP rs738409.⁴⁴ All the *PNPLA3* SNPs found in our study associated with HCC were not only in LD with rs738409, but also significant eQTL, suggesting their potential influences on protein function. Recent research suggests that *PNPLA3* translated from the risk alleles for NAFLD are more resistant to ubiquitylation and degradation, which allows the protein to accumulate on lipid droplet with increased binding to CGI-58 (comparative gene identification 58), suppressing the effect of CGI-58 on ATGL (adipose triglyceride lipase) activation. ATGL catalyzes lipolysis and lipophagy, and ATGL inhibition leads to lipid accumulation. Thus, *PNPLA3* risk alleles play a role in NAFLD which further progresses to HCC.^{45,46}

HCC-associated SNPs in *PNPLA3* (rs2281135, rs2896019, and rs4823173) were in moderate LD with rs3761472 and rs3827385 in *SAMM50*, and these SNPs were eQTL for both *PNPLA3* and *SAMM50* in skin and adipose tissues, respectively. SNPs in these genes are often described as one genomic block influencing the genetic susceptibility to NAFLD.^{34–36,47} Like *PNPLA3*, *SAMM50* is another protein regulating energy homeostasis. *SAMM50* is localized on the mitochondrial outer membrane and is involved in the maintenance of mitochondrial structure and morphology which plays a role in regulation of mitochondrial function.^{48–50} Recent experiments suggest that *SAMM50* is a key molecule in keeping mitochondria from mitophagy.⁵¹ While mitophagy is an important mechanism in maintaining the integrity of mitochondria, excessive mitophagy can lead to significant disruption of energy homeostasis, fatty acid synthesis, innate immunity, and apoptosis.⁵² Even though evidence linking the genotypes of *PNPLA3* and *SAMM50* to NAFLD and HCC is compelling, the molecular mechanism underlying the involvement of these proteins in the pathogenesis of HCC remains poorly understood. More studies are needed to elucidate their biological mechanisms in liver cancer.

Previously, we found that obesity and HCV were associated with HCC both independently and synergistically.¹⁶ The synergy was shown not only by a stronger association with the presence of both risk factors but also by younger ages at HCC diagnosis. HCC patients who were obese and infected with HCV were diagnosed at ages 9 years younger than those without these risk factors or 3–7 years younger than those with only one risk factor, suggesting an early onset of the disease due to heightened risk exposure. Since obesity can lead to NAFLD and NAFLD is associated with HCC, the association between HCC and *PNPLA3* risk alleles can be explained either by its connection to NAFLD only or through both NAFLD and HCV. As our GWAS has a relatively small sample size, we cannot really dissect these associations specifically and reliably. Obesity can lead not only to NAFLD, but also metabolic syndrome and type 2 diabetes. These health outcomes develop relatively sooner after obesity compared to obesity-related HCC which typically takes much longer time

to develop if no other major risk factors for HCC are present simultaneously. This difference in time course probably creates a phenomenon that HCC in areas where obesity is prevalent is also associated with type 2 diabetes and metabolic syndrome. Such associations which have been seen by us and others in earlier studies should be considered superficial, and the real causal connection is obesity. Compared to China, HBV is a much less significant risk factor for HCC in the United States.

Our study has several limitations. First, the sample size is relatively small, which limits the study power. Second, our study populations are not genetically homogenous. To minimize the influence of genetic inflation, we adjusted our analyses with the first three PCs. We also analyzed the data in whites only, and the results did not change substantially. Based on the 1000 Genomes data, the SNPs in *PNPLA3* and *SAMM50* that we found in association with HCC have high frequencies of alternative alleles across racial and ethnic groups (Table S6). This may help to alleviate the impact of population stratification because no racial/ethnic group loses the alternative alleles. Third, our de novo replication was done only in a small US study. Most of our replications were secondary data analysis on existing GWAS in Asia. Some of the GWAS focused on HCC either related to HCV or HBV, which made the replication complicated. Despite the inconsistency, our meta-analysis on the Asian studies still showed significant or borderline significant summary associations (Fig. S3), suggesting that SNPs in *PNPLA3/SAMM50* may have a role in HCC in both US and Asian populations.

In summary, our study showed that five SNPs in 22q13.31, three in *PNPLA3* (rs2281135, rs2896019, and 4823173) and two in *SAMM50* (rs3761472 and rs3827385), were associated with HCC. Meta-analysis of multiple datasets from Japan, Singapore and Taiwan confirmed that these SNPs were significantly associated with HCC, suggesting that genetic susceptibility to HCC may exist in 22q13.31.

References

- 1 Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J. Clin.* 2018; **68**: 394–424.
- 2 Ferlay J, Colombet M, Soerjomataram I *et al.* Estimating the global cancer incidence and mortality in 2018: GLOBOCAN sources and methods. *Int. J. Cancer.* 2019; **144**: 1941–53.
- 3 Ryerson AB, Ehemann CR, Altekruse SF *et al.* Annual Report to the Nation on the Status of Cancer, 1975–2012, featuring the increasing incidence of liver cancer. *Cancer.* 2016; **122**: 1312–37.
- 4 El-Serag HB, Kanwal F. Epidemiology of hepatocellular carcinoma in the United States: where are we? Where do we go? *Hepatology.* 2014; **60**: 1767–75.
- 5 Makarova-Rusher OV, Altekruse SF, McNeel TS *et al.* Population attributable fractions of risk factors for hepatocellular carcinoma in the United States. *Cancer.* 2016; **122**: 1757–65.
- 6 Miki D, Ochi H, Hayes CN *et al.* Variation in the *DEPDC5* locus is associated with progression to hepatocellular carcinoma in chronic hepatitis C virus carriers. *Nat. Genet.* 2011; **43**: 797–800.
- 7 Kumar V, Kato N, Urabe Y *et al.* Genome-wide association study identifies a susceptibility locus for HCV-induced hepatocellular carcinoma. *Nat. Genet.* 2011; **43**: 455–8.
- 8 Sawai H, Nishida N, Khor SS *et al.* Genome-wide association study identified new susceptible genetic variants in HLA class I region for

- hepatitis B virus-related hepatocellular carcinoma. *Sci. Rep.* 2018; **8**: 7958.
- 9 Zhang H, Zhai Y, Hu Z *et al.* Genome-wide association study identifies 1p36.22 as a new susceptibility locus for hepatocellular carcinoma in chronic hepatitis B virus carriers. *Nat. Genet.* 2010; **42**: 755–8.
 - 10 Li Y, Zhai Y, Song Q *et al.* Genome-wide association study identifies a new locus at 7q21.13 associated with hepatitis B virus-related hepatocellular carcinoma. *Clin. Cancer Res.* 2018; **24**: 906–15.
 - 11 Jiang DK, Sun J, Cao G *et al.* Genetic variants in STAT4 and HLA-DQ genes confer risk of hepatitis B virus-related hepatocellular carcinoma. *Nat. Genet.* 2013; **45**: 72–5.
 - 12 Li S, Qian J, Yang Y *et al.* GWAS identifies novel susceptibility loci on 6p21.32 and 21q21.3 for hepatocellular carcinoma in chronic hepatitis B virus carriers. *PLoS Genet.* 2012; **8**: e1002791.
 - 13 Chan KY, Wong CM, Kwan JS *et al.* Genome-wide association study of hepatocellular carcinoma in Southern Chinese patients with chronic hepatitis B virus infection. *PLoS One.* 2011; **6**: e28798.
 - 14 Lee MH, Huang YH, Chen HY *et al.* Human leukocyte antigen variants and risk of hepatocellular carcinoma modified by hepatitis C virus genotypes: a genome-wide association study. *Hepatology.* 2018; **67**: 651–61.
 - 15 Peng L, Yang G, Wu C, Wang W, Wu J, Guo Z. Mutations in hepatitis B virus small S genes predict postoperative survival in hepatocellular carcinoma. *Oncol. Targets. Ther.* 2016; **9**: 7367–72.
 - 16 Shen Y, Risch H, Lu L *et al.* Risk factors for hepatocellular carcinoma (HCC) in the northeast of the United States: results of a case-control study. *Cancer Causes Control.* 2020; **31**: 321–32.
 - 17 Yu WK, Shaw SM, Peck GE. Determination of dissolution profiles for several commercially available therapeutic [131I]sodium iodide capsules. *Int. J. Rad. Appl. Instrum. B.* 1990; **17**: 465–7.
 - 18 Goto A, Yamaji T, Sawada N *et al.* Diabetes and cancer risk: a Mendelian randomization study. *Int. J. Cancer.* 2020; **146**: 712–19.
 - 19 Tsugane S, Sawada N. The JPHC study: design and some findings on the typical Japanese diet. *Jpn. J. Clin. Oncol.* 2014; **44**: 777–82.
 - 20 Nagai A, Hirata M, Kamatani Y *et al.* Overview of the BioBank Japan project: study design and profile. *J. Epidemiol.* 2017; **27**: S2–8.
 - 21 Yuan JM, Stram DO, Arakawa K, Lee HP, Yu MC. Dietary cryptoxanthin and reduced risk of lung cancer: the Singapore Chinese Health Study. *Cancer Epidemiol. Biomarkers Prev.* 2003; **12**: 890–8.
 - 22 Chen CJ, Iloeje UH, Yang HI. Long-term outcomes in hepatitis B: the REVEAL-HBV study. *Clin. Liver Dis.* 2007; **11**: 797–816, viii.
 - 23 Purcell S, Neale B, Todd-Brown K *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* 2007; **81**: 559–75.
 - 24 Yang J, Lee SH, Goddard ME, Visscher PM. GCTA: a tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.* 2011; **88**: 76–82.
 - 25 Pruim RJ, Welch RP, Sanna S *et al.* LocusZoom: regional visualization of genome-wide association scan results. *Bioinformatics.* 2010; **26**: 2336–7.
 - 26 DerSimonian R, Laird N. Meta-analysis in clinical trials. *Control. Clin. Trials.* 1986; **7**: 177–88.
 - 27 Boyle AP, Hong EL, Hariharan M *et al.* Annotation of functional variation in personal genomes using RegulomeDB. *Genome Res.* 2012; **22**: 1790–7.
 - 28 Yang J, Trepo E, Nahon P *et al.* PNPLA3 and TM6SF2 variants as risk factors of hepatocellular carcinoma across various etiologies and severity of underlying liver diseases. *Int. J. Cancer.* 2019; **144**: 533–44.
 - 29 Singal AG, Manjunath H, Yopp AC *et al.* The effect of PNPLA3 on fibrosis progression and development of hepatocellular carcinoma: a meta-analysis. *Am. J. Gastroenterol.* 2014; **109**: 325–34.
 - 30 Li JF, Zheng EQ, Xie M. Association between rs738409 polymorphism in patatin-like phospholipase domain-containing protein 3 (PNPLA3) gene and hepatocellular carcinoma susceptibility: Evidence from case-control studies. *Gene.* 2019; **685**: 143–8.
 - 31 Kawaguchi T, Shima T, Mizuno M *et al.* Risk estimation model for nonalcoholic fatty liver disease in the Japanese using multiple genetic markers. *PLoS One.* 2018; **13**: e0185490.
 - 32 Romeo S, Kozlitina J, Xing C *et al.* Genetic variation in PNPLA3 confers susceptibility to nonalcoholic fatty liver disease. *Nat. Genet.* 2008; **40**: 1461–5.
 - 33 DiStefano JK, Kingsley C, Craig Wood G *et al.* Genome-wide analysis of hepatic lipid content in extreme obesity. *Acta Diabetol.* 2015; **52**: 373–82.
 - 34 Kitamoto T, Kitamoto A, Yoneda M *et al.* Genome-wide scan revealed that polymorphisms in the PNPLA3, SAMM50, and PARVB genes are associated with development and progression of non-alcoholic fatty liver disease in Japan. *Hum. Genet.* 2013; **132**: 783–92.
 - 35 Chung GE, Lee Y, Yim JY *et al.* Genetic polymorphisms of PNPLA3 and SAMM50 are associated with nonalcoholic fatty liver disease in a Korean population. *Gut Liver.* 2018; **12**: 316–23.
 - 36 Yuan X, Waterworth D, Perry JR *et al.* Population-based genome-wide association studies reveal six loci influencing plasma levels of liver enzymes. *Am. J. Hum. Genet.* 2008; **83**: 520–8.
 - 37 Li Q, Qu HQ, Rentfro AR *et al.* PNPLA3 polymorphisms and liver aminotransferase levels in a Mexican American population. *Clin. Invest. Med.* 2012; **35**: E237–45.
 - 38 Chen L, Lin Z, Jiang M *et al.* Genetic variants in the SAMM50 gene create susceptibility to nonalcoholic fatty liver disease in a Chinese Han population. *Hepat. Mon.* 2015; **15**: e31076.
 - 39 Yang A, Mottillo EP. Adipocyte lipolysis: from molecular mechanisms of regulation to disease and therapeutics. *Biochem. J.* 2020; **477**: 985–1008.
 - 40 Dubuquoy C, Robichon C, Lasnier F *et al.* Distinct regulation of adiponutrin/PNPLA3 gene expression by the transcription factors ChREBP and SREBP1c in mouse and human hepatocytes. *J. Hepatol.* 2011; **55**: 145–53.
 - 41 Baulande S, Lasnier F, Lucas M, Pairault J. Adiponutrin, a transmembrane protein corresponding to a novel dietary- and obesity-linked mRNA specifically expressed in the adipose lineage. *J. Biol. Chem.* 2001; **276**: 33336–44.
 - 42 Kershaw EE, Hamm JK, Verhagen LA, Peroni O, Katic M, Flier JS. Adipose triglyceride lipase: function, regulation by insulin, and comparison with adiponutrin. *Diabetes.* 2006; **55**: 148–57.
 - 43 Pirazzi C, Adiels M, Burza MA *et al.* Patatin-like phospholipase domain-containing 3 (PNPLA3) I148M (rs738409) affects hepatic VLDL secretion in humans and in vitro. *J. Hepatol.* 2012; **57**: 1276–82.
 - 44 Yang A, Mottillo EP, Mladenovic-Lucas L, Zhou L, Granneman JG. Dynamic interactions of ABHD5 with PNPLA3 regulate triacylglycerol metabolism in brown adipocytes. *Nat. Metab.* 2019; **1**: 560–9.
 - 45 Shang L, Mashek DG. The underpinnings of PNPLA3-mediated fatty liver emerge. *Hepatology.* 2020; **71**: 375–7.
 - 46 BasuRay S, Wang Y, Smagris E, Cohen JC, Hobbs HH. Accumulation of PNPLA3 on lipid droplets is the basis of associated hepatic steatosis. *Proc. Natl. Acad. Sci. USA.* 2019; **116**: 9521–6.
 - 47 Namjou B, Lingren T, Huang Y *et al.* GWAS and enrichment analyses of non-alcoholic fatty liver disease identify new trait-associated genes and pathways across eMERGE Network. *BMC Med.* 2019; **17**: 135.
 - 48 Sastri M, Darshi M, Mackey M *et al.* Sub-mitochondrial localization of the genetic-tagged mitochondrial intermembrane space-bridging components Mic19, Mic60 and Sam50. *J. Cell Sci.* 2017; **130**: 3248–60.
 - 49 Ding C, Wu Z, Huang L *et al.* Mitofilin and CHCHD6 physically interact with Sam50 to sustain cristae structure. *Sci. Rep.* 2015; **5**: 16064.
 - 50 Utsumi T, Matsuzaki K, Kiwado A *et al.* Identification and characterization of protein N-myristoylation occurring on four human mitochondrial proteins, SAMM50, TOMM40, MIC19, and MIC25. *PLoS One.* 2018; **13**: e0206355.

- 51 Jian F, Chen D, Chen L *et al.* Sam50 regulates PINK1-Parkin-mediated mitophagy by controlling PINK1 stability and mitochondrial morphology. *Cell Rep.* 2018; **23**: 2989–3005.
- 52 Pickles S, Vigie P, Youle RJ. Mitophagy and quality control mechanisms in mitochondrial maintenance. *Curr. Biol.* 2018; **28**: R170–85.

Supporting information

Additional supporting information may be found in the online version of this article at the publisher's website:

Supplemental Figure S1. Principle component analysis of the GWAS data. Panel A: PC1 versus PC2. Panel B: PC1 versus PC3. Panel C: PC2 versus PC3. Factor (race): red = white, green = black, blue = others, gray = not available.

Supplemental Figure S2. Levels of mRNA expression of *PNPLA3* and *SAMM50* in various tissues and organs analyzed by the Human Protein Atlas (<https://www.proteinatlas.org/>).

Supplemental Figure S3. Results of meta-analysis on the Asian studies with genotype data on the five SNPs in 22q13.31. Six studies had genotype data available for rs2281135, and four studies had data for rs2896019, rs3761472, rs3827385, and rs4823173.

Supplemental Table S1. Features of Replication Studies.

Supplemental Table S2. Linkage disequilibrium (r^2) between SNPs in 22q13.31.

Supplemental Table S3. Associations between HCC and top SNPs in GWAS stratified by viral infection.

Supplemental Table S4. Prediction of Potential Function of the SNPs in 22q13.31.

Supplemental Table S5. eQTL of the SNPs in 22q13.31 by GTEx.

Supplemental Table S6. SNP Allele Frequency in Different Population.