Genome-Wide Association Study of Pancreatic Cancer in Japanese Population

Siew-Kee Low[1,2]⁹, Aya Kuchiba[3]⁹, Hitoshi Zembutsu[1], Akira Saito[3,6], Atsushi Takahashi[4], Michiaki Kubo[5], Yataro Daigo[1,2], Naoyuki Kamatani[4], Suenori Chiku[3,7], Hirohiko Totsuka[3,8], Sumiko Ohnami[3], Hiroshi Hirose[9], Kazuaki Shimada[10], Takuji Okusaka[11], Teruhiko Yoshida[3]*, Yusuke Nakamura[1]*, Hiromi Sakamoto[3]

[1] Laboratory of Molecular Medicine, Human Genome Center, Institute of Medical Science, the University of Tokyo, Tokyo, Japan, [2] Department of Medical Genome Sciences, Graduate School of Frontier Sciences, the University of Tokyo, Tokyo, Japan, [3] Genetics Division, National Cancer Center Research Institute, Tokyo, Japan, [4] Laboratory for Statistical Analysis, Center for Genomic Medicine, RIKEN, Tokyo, Japan, [5] Laboratory for Genotyping Development, Center for Genomic Medicine, RIKEN, Kanagawa, Japan, [6] Statistical Genetics Analysis Division, StaGen Co., Ltd., Tokyo, Japan, [7] Science Solutions Division, Mizuho Information and Research Institute, Inc., Tokyo, Japan, [8] Bioinfomatics Group, Research and Development Center, Hitachi Government and Public Corporation System Engineering Ltd., Tokyo, Japan, [9] Department of Internal Medicine, Keio University School of Medicine, Tokyo, Japan, [10] Hepatobiliary and Pancreatic Surgery Division, National Cancer Center Hospital, Tokyo, Japan, [11] Hepatobiliary and Pancreatic Oncology Division, National Cancer Center Hospital, Tokyo, Japan

Abstract

Pancreatic cancer shows very poor prognosis and is the fifth leading cause of cancer death in Japan. Previous studies indicated some genetic factors contributing to the development and progression of pancreatic cancer; however, there are limited reports for common genetic variants to be associated with this disease, especially in the Asian population. We have conducted a genome-wide association study (GWAS) using 991 invasive pancreatic ductal adenocarcinoma cases and 5,209 controls, and identified three loci showing significant association (P-value$<5\times10^{-7}$) with susceptibility to pancreatic cancer. The SNPs that showed significant association carried estimated odds ratios of 1.29, 1.32, and 3.73 with 95% confidence intervals of 1.17–1.43, 1.19–1.47, and 2.24–6.21; P-value of 3.30×10^{-7}, 3.30×10^{-7}, and 4.41×10^{-7}; located on chromosomes 6p25.3, 12p11.21 and 7q36.2, respectively. These associated SNPs are located within linkage disequilibrium blocks containing genes that have been implicated some roles in the oncogenesis of pancreatic cancer.

Citation: Low S-K, Kuchiba A, Zembutsu H, Saito A, Takahashi A, et al. (2010) Genome-Wide Association Study of Pancreatic Cancer in Japanese Population. PLoS ONE 5(7): e11824. doi:10.1371/journal.pone.0011824

Editor: Wenjun Li, Duke University Medical Center, United States of America

Received May 7, 2010; **Accepted** July 3, 2010; **Published** July 29, 2010

Funding: This work was conducted as a part of Biobank Japan Project that was supported by the Ministry of Education, Culture, Sports, Sciences and Technology from the Japanese Government and Center for Genomic Medicine, RIKEN, Japan. It was also supported in part by the program for promotion of Fundamental Studies in Health Sciences of the National Institute of Biomedical Innovation (NiBio). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: Although the affiliations of the following three authors (Akira Saito [Statistical Genetics Analysis Division, StaGen Co., Ltd.], Suenori Chiku [Science Solutions Division, Mizuho Information and Research Institute, Inc.] and Hirohiko Totsuka [Bioinfomatics Group, Research and Development Center, Solution Division]) are in the private sector, each of them worked for one of the corresponding authors, Teruhiko Yoshida, as a part of a contract research solely funded by an academic research fund granted from NiBio (http://www.nibio.go.jp/english/) to T. Yoshida. Therefore, this study neither received any funding from the companies to which the above three authors belong, nor did their affiliating companies per se play any role in this research except that they sent their staff (the above three authors) to the laboratory of Teruhiko Yoshida under a research contract paid by Yoshida's grant. There is no competing interests involved between Teruhiko Yoshida and these companies, including intellectual properties. The authors also confirm that they can adhere to all the PLoS ONE policies on sharing data and materials.

* E-mail: yusuke@ims.u-tokyo.ac.jp (YN); tyoshida@ncc.go.jp (TY)

⁹ These authors contributed equally to this work.

Introduction

Pancreatic cancer is the fifth leading cause of cancer death with an estimated death of 24,634 patients in Japan in year 2007. Its 5-year survival rate is as low as 6.7% (http://www.fpcr.or.jp/publication/pdf/statistics2009/fig01.pdf and http://www.fpcr.or.jp/publication/pdf/statistics2009/fig20.pdf). Since no specific symptom is observed in the patients with pancreatic cancer at an early stage, most of the patients were diagnosed at their advanced stage with a very low possibility of cure for the disease [1,2].

Previous reports indicated the involvement of both environmental and genetics factors in the etiology of this deleterious disease. Several case-control and cohort epidemiological studies have identified a number of possible risk factors such as smoking [3], diabetes [4], chronic pancreatitis [5], which are likely to predispose individual to the disease. In addition, familial aggregation of the disease has implied the possible involvement of genetic factors in pancreatic cancer [6]; approximately 10% of the patients were reported to have family history and individuals having first-degree relatives with pancreatic cancer revealed 2- to 4- fold higher risk of the disease [7–9]. These data indicated that genetic factors are likely to play some roles in the development of pancreatic cancer. In the last decade, the advancement of molecular biology improved the understanding of the pathogenesis of pancreatic cancer and characterized a number of genes that mutated in pancreatic cancers, such as somatic mutations in genes *INK4A(CDKN2A), TP53, DPC4, BRCA1/2, STK11, APC, KRAS* and *ATM and PALB2* are found in pancreatic cancers [10–18].

Two recent GWAS studies for pancreatic cancer using Caucasian populations have identified associations with genome-wide significance on chromosomes 9p34.2 (*ABO*), 13q22.1, 1q32 (*NR5A2*) and 5p15.33 (*CLPTM1L-TERT*), and highlighted that accumulation of these common genetic risk variants with modest effects are likely to play an important role on this complex disease, either individually or in interaction with environmental factors [19–22]. As the ethnicity is one of the critical factors in the pathogenesis of the genetic diseases with complex gene-gene and gene-environmental interactions, we (Biobank Japan (BBJ) in The University of Tokyo and National Cancer Center (NCC) Japan) combined samples of 991 cases with pancreatic cancer and 5209 controls (Table S1), attempted to identify common genetic variations associated with susceptibility to pancreatic cancer in the Japanese population.

Results

After the standard quality control of the genotype results (Table S2), association analysis was performed for 420,236 SNPs using logistic regression analysis on the basis of allelic, dominant and recessive models after adjustment of age, sex and smoking status for each individual. The Q-Q plot for this GWAS based on allelic *P*-values by logistic regression revealed no significant population stratification with genomic inflation factor λ of 1.026 (Figure 1).

We successfully identified three genomic regions, 6p25.3, 12p11.21 and 7q36.2, shown to be significantly associated (*P*-value$<5.0\times10^{-7}$) with increased risk of pancreatic cancer in Japanese population as indicated in the Manhattan plot in Figure 2 (referred to ref. 23).

The most significantly-associated SNP, rs9502893 (*P*-value of 3.30×10^{-7}, per-allele odds ratio (OR) of 1.29 with 95% confidence interval (CI) of 1.17–1.43), is located within a 75-kb linkage disequilibrium (LD) block on chromosome 6p25.3 (Table 1). This LD block includes *FOXQ1* (forkhead box (Fox) Q1) gene, which is located 25 kb upstream to this marker SNP (Figure 3a). Imputation analysis also revealed modest association at SNPs located near to or on the *FOXQ1* gene suggesting it to be one of the causative genes for pancreatic cancer (Figure 3a and Table S3).

The second significantly-associated SNP, rs708224, located in the second intron of the gene *BICD1* (Bicaudal-D homolog 1) on chromosome 12p11 (*P*-value of 3.30×10^{-7}, per-allele OR of 1.32 with 95% CI of 1.19–1.47) (Table 1). The 80-kb LD block showing the association corresponds to the second intron of *BICD1* as revealed by the imputation analysis shown in Figure 3b (Table S3).

The third locus is marked by rs6464375, rs7779540, rs6973850 and rs1048768 in the first intron of *DPP6* gene. These SNPs indicated suggestive associations only under recessive model with minimum *P*-value of 4.41×10^{-7} (OR of 3.73 with 95%CI of 2.24–6.21) as shown in Table 1 and Figure 3c.

Discussion

Here we present results of GWAS analysis on 991 cases with pancreatic cancer and 5209 controls. Our study represents the first GWAS attempt to identify common variants associated with pancreatic cancer in Japanese population and successfully identified SNPs located on chromosomal loci of 6p25.3, 12p11.21 and 7q36.2 are significantly associated with increased risk of pancreatic cancer in Japanese population.

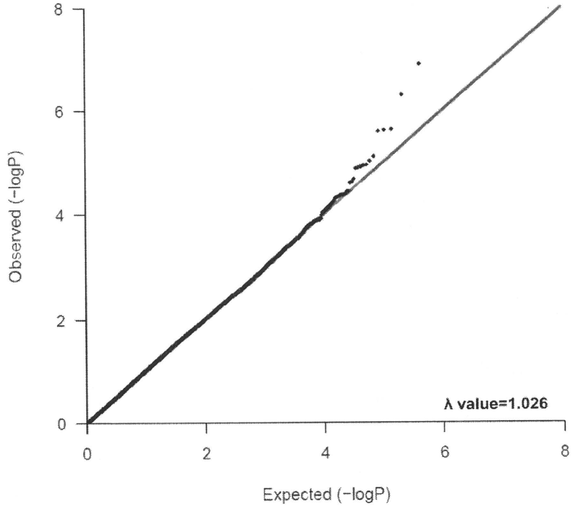

Figure 1. Q-Q plot for GWAS of pancreatic cancer in Japanese population. This Q-Q plot is based on logistic regression allelic *P*-values after standard quality control. (genomic inflation factor λ = 1.026).

doi:10.1371/journal.pone.0011824.g001

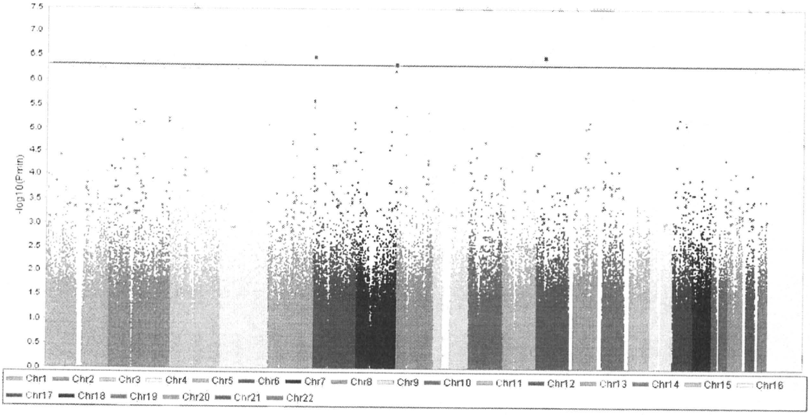

Figure 2. Manhattan plot for GWAS of pancreatic cancer in Japanese population. The plot is based on logistic regression model after correction of age, sex and smoking status. The P_{min} indicates the minimum *P*-value from logistic regression analysis for three models: allelic, dominant and recessive. Red line indicates genome-wide significant level (*P*-value $= 5 \times 10^{-7}$).
doi:10.1371/journal.pone.0011824.g002

It is known that the development of the common disease is caused by the accumulation of common genetic variants, and each of this variant has a very modest effect on the risk (for example OR of <1.2). In order to detect such small fraction, GWAS involving much larger populations (5000–10000) should be required. Our study was expected to identify SNPs with moderate effects (i.e OR>1.4). Hence SNPs that show very modest effect might have failed to be identified through this study.

The most significantly associated SNP in this GWAS, rs9502893 (*P*-value $= 3.30 \times 10^{-7}$, OR = 1.29) is located within a 75 kb LD block which encompasses gene *FOXQ1* on chromosome loci 6p25.3. *FOXQ1* encodes for protein forkhead box (Fox) Q1. The Fox family of transcription factors consists of at least 43 members and mutations in Fox genes can cause significant effects on human common disease and cancers [24,25]. A Fox member, FoxM1, is well-known to be associated with oncogenesis of pancreatic cancer. Down-regulation of this protein results in the inhibition of migration, invasion and angiogenesis in pancreatic cancer cells [26]. Furthermore, a recent study showed that FoxQ1 is overexpressed in pancreatic cancer, suggesting its role in pancreatic cancer tumorigenesis [27]. Although the SNP that we identified is approximately 25 kb downstream to this gene, the associated SNP may 'tag' the causative variant located on the expression regulatory region of the gene and subsequently alter expression of the gene. However, further study is needed to elucidate a precise biological role and mechanism of the gene function with regard to pancreatic carcinogenesis.

The second most significantly associated SNP, rs708224 (*P*-value $= 3.30 \times 10^{-7}$, OR = 1.32) is located within the *BICD1* gene. This gene encodes a protein Bicaudal-D homolog 1, which plays a role in vacuolar trafficking. Previous studies reported substantial evidences indicating a link between vacuolar gene and shorter telomeres in yeast model [28–30]. In addition, Mangino et al. suggested that genetic variations within the *BICD1* gene could alter its transcriptional levels and in turn influence telomere length in humans [31]. Several recent studies have documented reduced telomere length in pancreatic ductal adenocarnoma specimens, suggesting telomeric dysfunction in pancreatic cancer cells [32–34]. Thus, it is of importance to determine the functional consequences of rs708224 and/or variations linked to this SNP in the pathogenesis of pancreatic cancer.

Several SNPs located in the first intron of *DPP6* indicated suggestive associations with an increased risk of pancreatic cancer in this study. *DPP6* encodes protein dipeptidyl-peptidase 6, which binds to specific voltage-gated potassium channels and alters their expression and biophysical properties. A recent study on core signaling pathways in human pancreatic cancers found three somatic mutations in DPP6 among 24 pancreatic cancer samples examined by detailed sequence analyses. This report also suggested that DPP6 might play a crucial role in regulation of invasion of pancreatic cancer cells [35]. Hence, our study strengthens the risk of DPP6 in pancreatic cancer and warrants further screening on this gene to confirm its association with pancreatic cancer.

Recent GWAS reports have indicated several loci on chromosomes 9p34.2, 13q22.1, 1q32.1 and 5p15.33 to be associated with an increased risk of pancreatic cancer in Caucasian population [21,22]. Among the significantly associated SNPs, rs9543325 on chromosome 13q22.1 showed moderate association in our study populations (*P*-value (allelic model) of 1.69×10^{-4}; OR of 1.21 with 95%CI of 1.10–1.34) (Table S4). On the other hand, SNPs on chromosomes 9p34.2 (rs505922) and 1q32.1 (rs3790844) showed a weak association in our study populations (*P*-values of 3.69×10^{-2} and 1.24×10^{-2}; ORs of 1.11 and 1.14 with 95% CI of 1.01–1.22 and 1.03–1.27, respectively) (Table S4). We were unable to replicate the remaining loci (*SHH* and two loci on chromosomes 5p15.33 and 15q14) in these reports, probably because most of these associated SNPs are either non-polymorphic or possess very low allelic frequencies (MAF = 0.01) in Japanese population. The power of our study was not sufficient enough to detect positive associations for

Table 1. SNPs that show suggestive association with increase risk of pancreatic cancer in Japanese population.

CHR*	SNP	Position*	Risk allele	RAF		Allelic				Dominant				Recessive				P_{min}	Gene	Relativeloc*
				Case	Control	*P*-value	OR	L95	U95	*P*-value	OR	L95	U95	*P*-value	OR	L95	U95			
6	rs9502893	1285189	G	0.411	0.351	3.30E-07	1.29	1.17	1.43	2.97E-05	1.36	1.18	1.57	2.18E-05	1.50	1.24	1.80	3.30E-07	*FOXQ1*	25196
12	rs708224	32327676	A	0.718	0.656	3.30E-07	1.32	1.19	1.47	8.54E-07	1.42	1.23	1.63	2.09E-03	1.46	1.15	1.86	3.30E-07	*BICD1*	0
7	rs6464375	153256776	A	0.116	0.103	1.15E-01	1.13	0.97	1.32	7.36E-01	1.03	0.87	1.22	4.41E-07	3.73	2.24	6.21	4.41E-07	*DPP6*	0
7	rs7779540	153253595	A	0.116	0.103	1.08E-01	1.14	0.97	1.33	7.12E-01	1.03	0.87	1.23	4.58E-07	3.72	2.23	6.20	4.58E-07	*DPP6*	0
7	rs6973850	153269181	A	0.116	0.106	2.23E-01	1.10	0.94	1.29	9.76E-01	1.00	0.84	1.18	6.27E-07	3.64	2.19	6.04	6.27E-07	*DPP6*	0
6	rs11242679	1282311	A	0.366	0.311	2.40E-06	1.28	1.15	1.42	1.15E-05	1.37	1.19	1.58	2.07E-03	1.39	1.13	1.71	2.40E-06	*FOXQ1*	22318
6	rs7750826	1281867	G	0.365	0.311	2.57E-06	1.28	1.15	1.41	1.30E-05	1.37	1.19	1.57	1.98E-03	1.39	1.13	1.71	2.57E-06	*FOXQ1*	21874
7	rs10487687	153271407	A	0.150	0.136	8.99E-02	1.13	0.98	1.30	6.76E-01	1.03	0.88	1.21	3.35E-06	2.66	1.76	4.02	3.35E-06	*DPP6*	0
6	rs11242674	1252846	A	0.355	0.301	3.46E-06	1.28	1.15	1.41	9.64E-06	1.37	1.19	1.58	4.59E-03	1.37	1.10	1.69	3.46E-06	*FOXQ1*	−4829
2	rs6711606	101288602	A	0.135	0.116	1.27E-02	1.20	1.04	1.39	1.86E-01	1.12	0.95	1.32	4.02E-06	2.81	1.81	4.37	4.02E-06	*RNF149*	0
8	rs10088262	124834883	A	0.374	0.341	3.42E-03	1.16	1.05	1.28	4.30E-06	1.40	1.21	1.61	3.98E-01	0.91	0.74	1.13	4.30E-06	*FAM91A1*	−15180
8	rs7832232	38588460	A	0.483	0.454	1.43E-02	1.13	1.03	1.25	7.63E-01	0.98	0.84	1.14	5.10E-06	1.45	1.24	1.71	5.10E-06	*RNF5P1*	−10528
2	rs6736997	235279936	A	0.372	0.328	2.95E-04	1.20	1.09	1.33	4.96E-02	1.15	1.00	1.33	5.85E-06	1.57	1.29	1.91	5.85E-06	*ARL4C*	−209504
17	rs225190	27901771	G	0.410	0.360	5.99E-06	1.26	1.14	1.39	1.92E-04	1.32	1.14	1.52	2.37E-04	1.43	1.18	1.72	5.99E-06	*MYO1D*	0
2	rs4663158	235263691	A	0.397	0.352	1.39E-04	1.21	1.10	1.34	2.89E-02	1.17	1.02	1.35	6.91E-06	1.53	1.27	1.85	6.91E-06	*ARL4C*	−193259
13	rs2039553	79197723	A	0.291	0.268	1.32E-02	1.15	1.03	1.28	4.39E-01	1.06	0.92	1.21	7.01E-06	1.73	1.36	2.19	7.01E-06	*NDFIP2*	171501
2	rs1427593	137271694	A	0.110	0.080	1.55E-05	1.42	1.21	1.66	7.10E-06	1.49	1.25	1.77	4.30E-01	1.31	0.67	2.58	7.10E-06	*THSD7B*	−193238
6	rs3016539	162156065	A	0.903	0.871	1.67E-05	1.42	1.21	1.67	7.28E-06	1.50	1.26	1.79	2.90E-01	1.36	0.77	2.43	7.28E-06	*PARK2*	0
2	rs12615966	104745389	A	0.112	0.097	1.40E-02	1.22	1.04	1.42	1.59E-01	1.13	0.95	1.35	7.44E-06	3.15	1.91	5.21	7.44E-06	*LOC284998*	−6744
17	rs2257205	53803296	A	0.378	0.327	1.58E-05	1.25	1.13	1.38	7.74E-06	1.38	1.20	1.59	2.97E-02	1.25	1.02	1.53	7.74E-06	*RNF43*	0
5	rs6879627	2162901	G	0.575	0.522	8.12E-06	1.25	1.14	1.39	4.66E-04	1.31	1.12	1.52	1.57E-04	1.42	1.18	1.69	8.12E-06	*LOC731559*	225138
17	rs4924935	18694595	G	0.269	0.228	8.80E-05	1.25	1.12	1.40	8.15E-06	1.37	1.19	1.58	5.06E-01	1.11	0.82	1.48	8.15E-06	PRPSAP2	(7622
17	rs1737947	18772157	G	0.252	0.212	3.88E-05	1.27	1.13	1.43	8.49E-06	1.37	1.19	1.58	2.89E-01	1.19	0.87	1.62	8.49E-06	PRPSAP2	0
13	rs1886449	72830115	A	0.424	0.383	2.61E-04	1.21	1.09	1.33	5.62E-02	1.15	1.00	1.33	9.24E-06	1.51	1.26	1.80	9.24E-06	*LOC730242*	−206271
13	rs1585440	65379816	C	0.761	0.713	9.28E-06	1.30	1.16	1.45	3.09E-05	1.35	1.17	1.55	7.36E-03	1.50	1.12	2.03	9.28E-06	*LOC387933*	−118820
3	rs4683235	46476935	A	0.215	0.173	9.93E-06	1.31	1.16	1.48	7.53E-05	1.34	1.16	1.54	2.24E-03	1.70	1.21	2.38	9.93E-06	*LTF*	0

Odds ratios, 95% confidence limits and P-values were obtained using logistic regression analysis according to allelic, dominant and recessive model after adjustment of age, sex and smoking.
RAF, risk allele frequency; OR, odds ratio; L95, U95, lower and upper confidence limits; P_{min}, minimum *P*-value among three genetic models.
*Position and relative loci (Relativeloc) are based on NCBI Human Genome Build 36.
doi:10.1371/journal.pone.0011824.t001

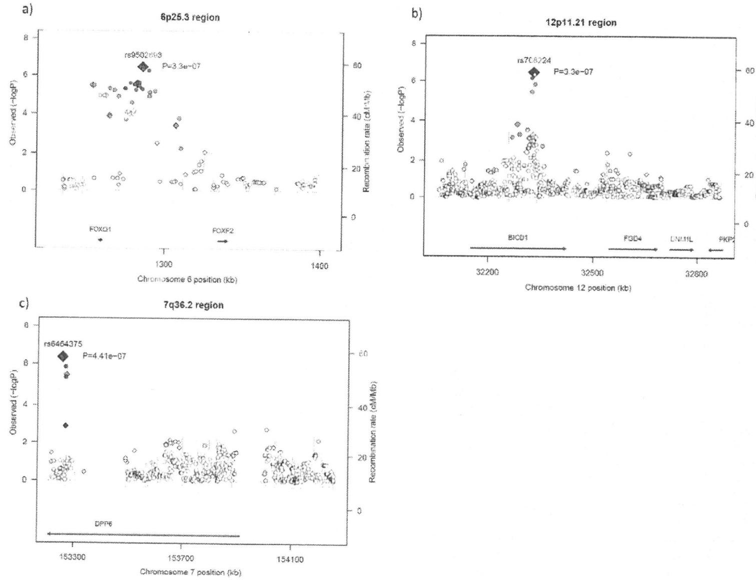

Figure 3. Regional association plots for three pancreatic cancer risk loci. (a) 6p25.3 region, SNP rs9502893 located 25 kb upstream to gene *FOXQ1*. (b) 12p11.21 region, SNP rs708224 is located at the second intron of gene *BICD1*. (c) 7q36.2 region, SNP rs6464375 is located at the first intron of gene *DPP6 transcript variant 3*. Each of the marker SNPs is marked by a blue diamond. SNPs that are genotyped in the Illumina platform are plotted as diamonds; Imputed SNPs are plotted as circles. The color intensity reflects the extent of LD with the marker SNP, red ($r^2 \geq 0.8$), orange ($0.5 \leq r^2 < 0.8$), yellow ($0.2 \leq r^2 < 0.5$) and white ($r^2 < 0.2$). Light blue line indicated local recombination rate.
doi:10.1371/journal.pone.0011824.g003

these variants with the low allelic frequency. Such ethnic difference in genetic architecture of disease susceptibility is not rare. For example, two recent GWAS reported common variants on *KCNQ1* gene associated with type 2 diabetes mellitus in Japanese population, but European GWAS were unable to identify the associations due to the low allelic frequency of these variants in the population [36,37]. In addition, identification of susceptibility loci may be also influenced by the differences in the LD structure across different populations and by potential interaction with other genetic variants and environmental factors [38].

In summary, this study represents the first GWAS to identify common variants possibly associated with pancreatic cancer in Japanese population. Our study confirmed the association from the Caucasian GWAS studies and revealed several novel possible candidate associated loci that were not detected in the previous Caucasian GWAS studies. Nevertheless, further additional replications are required to confirm or exclude the current findings.

Materials and Methods

Case and control subjects

A total of 331 and 675 cases that were clinically and/or histologically diagnosed to have an invasive pancreatic ductal adenocarcinoma were obtained from Biobank Japan (http://biobankjp.org) at the Institute of Medical Science, The University of Tokyo as well as National Cancer Center Hospital, respectively. The control samples consisted of Japanese volunteers that were obtained from Osaka-Midosuji Rotary Club, Osaka, Japan ($n = 906$) as well as from staff members in Keio University, Japan, who participated in its health-check program ($n = 677$). In addition, individuals who were registered in Biobank Japan as subjects with various diseases except cancer ($n = 3,728$) (those having pulmonary tuberculosis, chronic hepatitis-B, keroid, drug-induced skin rash, peripheral artery disease, arrhythmia, stroke and myocardial infarction) were used as controls. All samples were obtained after obtaining the written informed consent. This project was approved by the ethics committee at The Institute of Medical Sciences, The University of Tokyo, National Cancer Center and Keio University. Individuals who had clinical history of diabetes mellitus (a possible confounding factor for pancreatic cancer) were excluded from these control sets. For sample quality control, we excluded five cases with call rate<0.98. After performing principal component analysis, we excluded outliers of 10 cases and 102 controls, who did not belong to the major Japanese cluster (Hondo cluster) (Figure S1) [39]. We eventually performed the association study based on 991 cases and 5209 controls (Table S1). Power calculation showed that our study would have over 90% power to detect a per-allele OR of 1.4 or greater for an allele with 30% frequency at the genome-wide significance level ($\alpha = 5\times10^{-7}$).

SNP genotyping and quality control

All the individuals were genotyped using either Illumina Infinium HumanHap550v3 or Illumina Infinium Human610-Quad DNA Analysis Genotyping BeadChip. SNPs common in the two platforms were used for further analysis. We applied SNP quality control for all sets of samples as follows; SNP call rate should be >0.99 in both cases and controls, and *P*-value of Hardy-Weinberg equilibrium test should be $>1.0\times10^{-6}$ in controls. SNPs with minor allele frequency (MAF) of <0.01 in both case and control samples were excluded from the further analysis (Table S2).

Statistical analysis

We analyzed each SNP using logistic regression adjusted for age (continuous), sex and smoking status (current/former, never). *P*-values and OR with 95%CI were calculated for allelic, dominant and recessive models. We used the minimum *P*-values obtained from three models to evaluate the statistical significance of the association. All OR were reported with respect to the risk allele. All the statistical analyses were performed using R statistical environment version 2.9.0 (http://www.r-project.org/) or PLINK 1.06 (http://pngu.mgh.harvard.edu/~purcell/plink/). R statistical environment version 2.9.0 was employed to draw Q-Q plot and regional association plot.

Genotype Imputation

We performed genotype imputation analysis for each set of samples by utilizing a Hidden Markov model as programmed in MACH version 1.0 (http://www.sph.umich.edu/csg/abecasis/mach/index.html). To infer untyped and missing genotypes around the candidate chromosomal loci, we provided genotypes from our own samples together with haplotypes for reference samples (Japanese from Tokyo, JPT) from HapMap database (http://hapmap.ncbi.nlm.nih.gov/). SNPs with low genotyping rate (<99%), showing deviations from Hardy-Weinberg equilibrium ($<1.0\times10^{-6}$), or MAF (<0.01) were excluded from the analysis. MACH version 1.0 was used to estimate haplotypes, map crossover and error rates using 50 iterations of the Markov chain Monte Carlo algorithm. By utilizing the genotype information from the HapMap database, maximum likelihood genotypes were generated. For quality control, we retained imputed SNPs with the estimated r^2 of >0.3. We also picked up a total of 17 SNPs (*P*-value<0.001) to verify the association using Invader and TaqMan genotyping methods (data not shown).

Supporting Information

Table S1 Sample characteristic of this study.
Found at: doi:10.1371/journal.pone.0011824.s001 (0.02 MB XLS)

Table S2 Total number of SNPs excluded according to each quality control criteria.
Found at: doi:10.1371/journal.pone.0011824.s002 (0.02 MB XLS)

Table S3 Imputation analysis around significantly associated SNPs.
Found at: doi:10.1371/journal.pone.0011824.s003 (0.04 MB XLS)

Table S4 Association study of SNPs which shown to be significantly associated with increased risk of pancreatic cancer in Caucasian population in Japanese.
Found at: doi:10.1371/journal.pone.0011824.s004 (0.02 MB XLS)

Figure S1 Principal component analysis for GWAS of pancreatic cancer in Japanese population. a) Principal component analysis for GWAS of pancreatic cancer in Japanese population refer to four HapMap population control subjects including CEU indicates Caucasians from Utah; YRI, Nigerians from Yoruba; CHB, Han Chinese from Beijing and JPT, Japanese from Tokyo. b) Principal component analysis of study subjects referred only to Asian populations. We utilized samples from the homogenous case-control (Hondo) cluster.
Found at: doi:10.1371/journal.pone.0011824.s005 (9.43 MB TIF)

Acknowledgments

We express our heartfelt gratitude to all the patients who participate in this study. We would like to thank Dr Yoichiro Kamatani for his constructive comments and suggestions. Our thankfulness also goes to the member of The Rotary Club of Osaka-Midosuji District 2660 Rotary International in Japan for making this study possible. We thank Drs. Hideki Ueno, Masafumi Ikeda, Chigusa Morizane, Yoshihiro Sakamoto, Minoru Esaki, Tomoo Kosuge and Nobuyoshi Hiraoka for ascertainment of the patients and their clinico-pathological information at the National Cancer Center Hospital. We also would like to express our gratefulness to Miss Kumi Matsuda for her outstanding technical assistance.

Author Contributions

Conceived and designed the experiments: SKL AK HZ MK YD NK TY YN HS. Performed the experiments: SKL AK MK SO HS. Analyzed the data: SKL AK HZ AS AT MK NK SC HT TY YN HS. Contributed reagents/materials/analysis tools: SKL AK HZ AS AT MK NK HH KS TO TY YN. Wrote the paper: SKL AK HZ TY YN.

References

1. Kelsen DP, Portenoy R, Thaler H, Tao Y, Brennan M (1997) Pain as a predictor of outcome in patients with operable pancreatic carcinoma. Surgery 122: 53–59.
2. Catanzaro A, Richardson S, Veloso H, Isenberg GA, Wong RC, et al. (2003) Long-term follow-up of patients with clinically indeterminate suspicion of pancreatic cancer and normal EUS. Gastrointest Endosc 58: 836–840.
3. Anderson KE, Mack T, Silverman D (2006) Cancer of the pancreas. In: Schottenfeld D, Fraumeni JF, Jr., eds. Cancer Epidemiology and Prevention. New York: Oxford University Press. pp 721–762.
4. Stevens RJ, Roddam AW, Beral V (2007) Pancreatic cancer in type 1 and young-onset diabetes: systematic review and meta-analysis. Br J Cancer 96: 507–509.
5. Lowenfels AB, Maisonneuve P, Cavallini G, Ammann RW, Lankisch PG, et al. (1993) Pancreatitis and the risk of pancreatic cancer. International Pancreatitis Study Group. N Engl J Med 328: 1433–1437.
6. Del Chiaro M, Zerbi A, Falconi M, Bertacca L, Polese M, et al. (2007) Cancer risk among the relatives of patients with pancreatic ductal adenocarcinoma. Pancreatology 7: 459–469.
7. McWilliams RR, Rabe KG, Olswold C, De Andrade M, Petersen GM (2005) Risk of malignancy in first-degree relatives of patients with pancreatic carcinoma. Cancer 104: 388–394.
8. Fernandez E, La Vecchia C, D'Avanzo B, Negri E, Franceschi S (1994) Family history and the risk of liver, gallbladder, and pancreatic cancer. Cancer Epidemiol Biomarkers Prev 3: 209–212.
9. Tersmette AC, Petersen GM, Offerhaus GJ, Falatko FC, Brune KA, et al. (2001) Increased risk of incident pancreatic cancer among first-degree relatives of patients with familial pancreatic cancer. Clin Cancer Res 7: 738–744.
10. Yan L, McFaul C, Howes N, Leslie J, Lancaster G, et al. (2005) Molecular analysis to detect pancreatic ductal adenocarcinoma in high-risk groups. Gastroenterology 128: 2124–2130.
11. Caldas C, Hahn SA, da Costa LT, Redston MS, Schutte M, et al. (1994) Frequent somatic mutations and homozygous deletions of the p16 (MTS1) gene in pancreatic adenocarcinoma. Nat Genet 8: 27–32.
12. Barton CM, Staddon SL, Hughes CM, Hall PA, O'Sullivan C, et al. (1991) Abnormalities of the p53 tumour suppressor gene in human pancreatic cancer. Br J Cancer 64: 1076–1082.

13. Berrozpe G, Schaeffer J, Peinado MA, Real FX, Perucho M (1994) Comparative analysis of mutations in the p53 and K-ras genes in pancreatic cancer. Int J Cancer 58: 185–191.
14. Hahn SA, Schutte M, Hoque AT, Moskaluk CA, da Costa LT, et al. (1996) DPC4, a candidate tumor suppressor gene at human chromosome 18q21.1. Science 271: 350–3.
15. Hruban RH, Iacobuzio-Donahue C, Wilentz RE, Goggins M, Kern SE (2001) Molecular pathology of pancreatic cancer. Cancer J 7: 251–258.
16. Hahn SA, Greenhalf B, Ellis I, Sina-Frey M, Rieder H, et al. (2003) BRCA2 germline mutations in familial pancreatic carcinoma. J Natl Cancer Inst 95: 214–221.
17. Wong T, Howes N, Threadgold J, Smart HL, Lombard MG, et al. (2001) Molecular diagnosis of early pancreatic ductal adenocarcinoma in high-risk patients. Pancreatology 1: 486–509.
18. Jones S, Hruban RH, Kamiyama M, Borges M, Zhang X, et al. (2009) Exomic sequencing identifies PALB2 as a pancreatic cancer susceptibility gene. Science 324: 217.
19. MacLeod SL, Chowdhury P (2006) The genetics of nicotine dependence: relationship to pancreatic cancer. World J Gastroenterol 12: 7433–7439.
20. Milne RL, Greenhalf W, Murta-Nascimento C, Real FX, Malats N (2009) The inherited genetic component of sporadic pancreatic adenocarcinoma. Pancreatology 9: 206–214.
21. Amundadottir L, Kraft P, Stolzenberg-Solomon RZ, Fuchs CS, Petersen GM, et al. (2009) Genome-wide association study identifies variants in the ABO locus associated with susceptibility to pancreatic cancer. Nat Genet 41: 986–990.
22. Petersen GM, Amundadottir L, Fuchs CS, Kraft P, Stolzenberg-Solomon RZ, et al. (2010) A genome-wide association study identifies pancreatic cancer susceptibility loci on chromosomes 13q22.1, 1q32.1 and 5p15.33. Nat Genet 42: 224–228.
23. Wellcome Trust Case Control Consortium (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 447: 661–78.
24. Myatt SS, Lam EW (2007) The emerging roles of forkhead box (Fox) proteins in cancer. Nat Rev Cancer 7: 847–859.
25. Hannenhalli S, Kaestner KH (2009) The evolution of Fox genes and their role in development and disease. Nat Rev Genet 10: 233–240.
26. Wang Z, Banerjee S, Kong D, Li Y, Sarkar FH (2007) Down-regulation of Forkhead Box M1 transcription factor leads to the inhibition of invasion and angiogenesis of pancreatic cancer cells. Cancer Res 67: 8293–8300.
27. Cao D, Hustinx SR, Sui G, Bala P, Sato N, et al. (2004) Identification of novel highly expressed genes in pancreatic ductal adenocarcinomas through a bioinformatics analysis of expressed sequence tags. Cancer Biol Ther 3: 1081–1089.
28. Askree SH, Yehuda T, Smolikov S, Gurevich R, Hawk J, et al. (2004) A genome-wide screen for Saccharomyces cerevisiae deletion mutants that affect telomere length. Proc Natl Acad Sci U S A 101: 8658–8663.
29. Gatbonton T, Imbesi M, Nelson M, Akey JM, Ruderfer DM, et al. (2006) Telomere Length as a Quantitative Trait: Genome-Wide Survey and Genetic Mapping of Telomere Length-Control Genes in Yeast. PLoS Genet 2: e35. doi:10.1371/journal.pgen.0020035.
30. Rog O, Smolikov S, Krauskopf A, Kupiec M (2005) The yeast VPS genes affect telomere length regulation. Curr Genet 47: 18–28.
31. Mangino M, Brouilette S, Braund P, Tirmizi N, Vasa-Nicotera M, et al. (2008) A regulatory SNP of the BICD1 gene contributes to telomere length variation in humans. Hum Mol Genet 17: 2518–2523.
32. Büchler P, Conejo-Garcia JR, Lehmann G, Müller M, Emrich T, et al. (2001) Real-time quantitative PCR of telomerase mRNA is useful for the differentiation of benign and malignant pancreatic disorders. Pancreas 22: 331–340.
33. Kobitsu K, Tsutsumi M, Tsujiuchi T, Suzuki F, Kido A, et al. (1997) Shortened telomere length and increased telomerase activity in hamster pancreatic duct adenocarcinomas and cell lines. Mol Carcinog 18: 153–159.
34. van Heek NT, Meeker AK, Kern SE, Yeo CJ, Lillemoe KD, et al. (2002) Telomere shortening is nearly universal in pancreatic intraepithelial neoplasia. Am J Pathol 161: 1541–1547.
35. Jones S, Zhang X, Parsons DW, Lin JC, Leary RJ, et al. (2008) Core signaling pathways in human pancreatic cancers revealed by global genomic analyses. Science 321: 1801–1806.
36. Unoki H, Takahashi A, Kawaguchi T, Hara K, Horikoshi M, et al. (2008) SNPs in KCNQ1 are associated with susceptibility to type 2 diabetes in East Asian and European populations. Nat Genet 40: 1098–1102.
37. Yasuda K, Miyake K, Horikawa Y, Hara K, Osawa H, et al. (2008) Variants in KCNQ1 are associated with susceptibility to type 2 diabetes mellitus. Nat Genet 40: 1092–1097.
38. McCarthy MI, Abecasis GR, Cardon LR, Goldstein DB, Little J, et al. (2008) Genome-wide association studies for complex traits: consensus, uncertainty and challenges. Nat Rev Genet 9: 356–369.
39. Yamaguchi-Kabata Y, Nakazono K, Takahashi A, Saito S, Hosono N, et al. (2008) Japanese population structure, based on SNP genotypes from 7003 individuals compared to other ethnic groups: effects on population-based association studies. Am J Hum Genet 83: 445–456.

nature genetics

High-resolution characterization of a hepatocellular carcinoma genome

Yasushi Totoki[1], Kenji Tatsuno[2], Shogo Yamamoto[2], Yasuhito Arai[1], Fumie Hosoda[1], Shumpei Ishikawa[3], Shuichi Tsutsumi[2], Kohtaro Sonoda[2], Hirohiko Totsuka[4], Takuya Shirakihara[1], Hiromi Sakamoto[4], Linghua Wang[2], Hidenori Ojima[5], Kazuaki Shimada[6], Tomoo Kosuge[6], Takuji Okusaka[7], Kazuto Kato[8], Jun Kusuda[9], Teruhiko Yoshida[4], Hiroyuki Aburatani[2] & Tatsuhiro Shibata[1]

Hepatocellular carcinoma, one of the most common virus-associated cancers, is the third most frequent cause of cancer-related death worldwide[1]. By massively parallel sequencing[2] of a primary hepatitis C virus–positive hepatocellular carcinoma (36× coverage) and matched lymphocytes (>28× coverage) from the same individual, we identified more than 11,000 somatic substitutions of the tumor genome that showed predominance of T>C/A>G transition and a decrease of the T>C substitution on the transcribed strand, suggesting preferential DNA repair. Gene annotation enrichment analysis[3] of 63 validated non-synonymous substitutions revealed enrichment of phosphoproteins. We further validated 22 chromosomal rearrangements, generating four fusion transcripts that had altered transcriptional regulation (*BCORL1-ELF4*) or promoter activity. Whole-exome sequencing[4,5] at a higher sequence depth (>76× coverage) revealed a *TSC1* nonsense substitution in a subpopulation of the tumor cells. This first high-resolution characterization of a virus-associated cancer genome identified previously uncharacterized mutation patterns, intra-chromosomal rearrangements and fusion genes, as well as genetic heterogeneity within the tumor.

We sequenced short-insert (250 bp, on average) genomic libraries of a primary hepatitis C virus (HCV)–positive hepatocellular carcinoma (HCC) and lymphocytes from a Japanese male (**Supplementary Fig. 1**) using the Illumina GAIIx sequencer with 50-bp paired-end reads. After alignment to the human reference genome and removal of PCR duplications, we obtained high-quality nucleotide sequences covering 102.5 Gb of the tumor genome (35.9× coverage) and 80.2 Gb (28.1× coverage) of the lymphocyte genome (**Supplementary Table 1**). The sequenced reads covered 99.69% (tumor) and 99.79% (lymphocyte) of the human reference genome. We identified 3,023,587 germline variations in the lymphocyte genome, approximately 90% of which were found in the dbSNP database, and 2,939,032 nucleotide variations in the tumor genome (a proportion of the variation was lost as a result of chromosomal alterations in the tumor genome). Comparison of the tumor and lymphocyte genomes revealed 11,731 somatically acquired nucleotide changes in the tumor genome (**Table 1**).

The prevalence of somatic substitutions was significantly less in the genic (intronic, non-coding exon and coding exon) regions relative to the intergenic regions (**Fig. 1a**, left), which could be partially explained by negative selection of lethal mutations in the gene regions or by the existence of specific molecules responsible for the repair of transcribed regions[6]. There was no significant difference in the prevalence of somatic substitutions between those of non-coding and coding exons (**Fig. 1a**, left), whereas the prevalence of germline variation was significantly decreased in the coding exons (**Fig. 1a**, right). Additionally, the ratio of non-synonymous to synonymous somatic substitutions (63/18 = 3.5) in the tumor genome was significantly higher than that of germline variations (9,573/10,552 = 0.91; $P < 0.0001$) but was not significantly different from that expected by chance (3.36; $P = 0.91$). This result suggests that an increase in negative selection of somatic substitution on the coding exons is weaker than that of germline variation. An alternative, but not mutually exclusive, explanation is that positive selection, which benefits the survival of tumor cells, partially occurs on the coding exons. The distribution of somatic substitutions revealed the dominance of T>C/A>G and C>T/G>A transitions (**Fig. 1b**). Sequence context preference was evident in some nucleotide substitutions. The C>T transition occurred significantly at CpG sites (15%; $P < 0.0001$), whereas the T>C transition occurred frequently at ApT sites (40%; $P < 0.0001$) (**Supplementary Fig. 2**). Only the T>C/A>G transition was significantly ($P = 0.01$) lower in the coding exons relative to the intergenic

[1]Division of Cancer Genomics, National Cancer Center Research Institute, Chuo-ku, Tokyo, Japan. [2]Genome Science Division, Research Center for Advanced Science and Technology, University of Tokyo, Meguro-ku, Tokyo, Japan. [3]Department of Pathology, Graduate School of Medicine, University of Tokyo, Bunkyo-ku, Tokyo, Japan. [4]Division of Genetics, National Cancer Center Research Institute, Chuo-ku, Tokyo, Japan. [5]Division of Molecular Pathology, National Cancer Center Research Institute, Chuo-ku, Tokyo, Japan. [6]Hepatobiliary and Pancreatic Surgery Division, National Cancer Center Hospital, Chuo-ku, Tokyo, Japan. [7]Hepatobiliary and Pancreatic Oncology Division, National Cancer Center Hospital, Chuo-ku, Tokyo, Japan. [8]Institute for Research in Humanities, Graduate School of Biostudies, Institute for Integrated Cell-Material Sciences, Kyoto University, Kyoto, Japan. [9]National Institute of Biomedical Innovation, Ibaraki, Osaka, Japan. Correspondence should be addressed to T. Shibata (tashibat@ncc.go.jp).

Received 21 July 2010; accepted 14 March 2011; published online 17 April 2011; doi:10.1038/ng.804

Table 1 Somatically acquired alterations in a liver cancer genome

Type of change	Number	Percentage
Substitutions	11,731	100.0
Coding	81	0.7
Nonsense	1	<0.1
Missense	62	0.5
Synonymous	18	0.2
Non-coding	120	1.0
UTR	83	0.7
Pseudogene	23	0.2
ncRNA	19	0.2
Intronic	4,001	34.1
Splice site	2	<0.1
Other	3,999	34.1
Intergenic	7,529	64.2
Small insertions and deletions	670	100.0
Coding	7	1.0
Non-coding	9	1.3
UTR	8	1.2
Pseudogene	0	0.0
ncRNA	2	0.3
Intronic	249	37.2
Splice site	0	0.0
Other	249	37.2
Intergenic	405	60.4
Rearrangements	22	100.0
Intrachromosomal	21	95.5
Deletions	11	50.0
Inversions	9	40.9
Tandem duplications	1	4.5
Interchromosomal	1	4.5

In 'non-coding' categories, some mutations have been classified into two subgroups. Four substitutions were classified as both UTR and non-coding RNA. One substitution was classified as both a pseudogene and non-coding RNA. One indel was classified as both UTR and non-coding RNA. UTR, untranslated region; ncRNA, non-coding RNA.

regions (**Fig. 1c**), and the C>T/G>A transition was more frequent in the coding exons relative to the intronic and non-coding exon regions, partly due to the higher GC content of coding exons and the higher frequency of CpG methylation. There were fewer T>C transitions on the transcribed strands than on the untranscribed strands ($P < 0.0001$) (**Fig. 1d**), and we observed no statistically significant differences for other substitutions.

We detected 90 somatic substitutions in protein-coding regions, 81 (including 63 non-synonymous substitutions) of which were validated as somatic alterations by Sanger sequencing of both the tumor and lymphocyte genomes (**Tables 1,2** and **Supplementary Fig. 3**). Of the remaining nine substitutions, three could not be amplified by PCR, four could not be sequenced due to the surrounding repetitive sequences and two could not be validated, likely because they were located within highly homologous segmental duplications or processed pseudogene regions. We also found evidence for 670 small somatic insertions and deletions, and all seven that are located in protein-coding regions were validated (**Tables 1** and **2**, **Supplementary Fig. 13**). These somatic alterations included mutations of two well-known tumor suppressor genes for HCC (*TP53* and *AXIN1*) and five genes (*ADAM22*, *JAK2*, *KHDRBS2*, *NEK8* and *TRRAP*) that have been found to be mutated in other cancers[7]. Gene annotation enrichment analysis[3] of the non-synonymous somatic mutations revealed significant overrepresentation of genes encoding phosphorproteins ($P = 0.0017$) and those with bipartite nuclear localization signals ($P = 0.029$) (**Supplementary Table 2**). Further re-sequencing of the exons containing potentially deleterious mutations in 96 additional pairs of primary HCC and non-cancerous liver and 21 HCC cell lines revealed two mutations (resulting in p.Phe190Leu and p.Gln212X, of which only the latter was proven to be somatic) in *LRRC30* (**Supplementary Fig. 4**). *LRRC30* contains nine repeats of a leucine-rich domain of unknown function, and all validated mutations changed the well-conserved amino acid in these repeats or produced a truncated protein.

We predicted 33 somatic rearrangements, 22 of which were validated by Sanger sequencing of the breakpoints in both the tumor and lymphocyte genomes (**Table 3**). Most of the rearrangements were intrachromosomal and occurred at the boundaries of copy number change (**Supplementary Fig. 5**). In particular, nine structural aberrations were clustered in the region of 11q12.2–11q13.4, generating a complex pattern of chromosomal amplification and loss (**Supplementary Fig. 6**). RT-PCR and sequencing analysis of the tumor and matched non-cancerous liver tissue validated four somatic fusion transcripts generated by rearrangements: the *BCORL1-ELF4* and *CTNND1-STX5* fusion genes by intra-chromosomal inversions (Xq25 and 11q12, respectively), the *VCL-ADK* fusion gene by an interstitial deletion in 10q22 (**Supplementary Fig. 7**) and the *CABP2-LOC645332* fusion gene by a tandem duplication in 11q13 (**Supplementary Fig. 8**). The *BCORL1-ELF4* chimeric transcript combining exons 1–11 of *BCORL1* and exon 8 of *ELF4* encodes an in-frame fusion protein (**Fig. 2a,b**). Quantitative RT-PCR revealed increased (>sixfold) expression of fusion transcripts in the tumor relative to wild-type *BCORL1* and *ELF4* gene expression in the non-cancerous liver (data not shown). *BCORL1* associates with CtBP and class II histone deacetylases and functions as a transcriptional repressor[8], and *ELF4* encodes a transcriptional activator[9,10] (**Fig. 2b**). We expressed *BCORL1*, *ELF4* and the chimera *BCORL1-ELF4* as Gal4-DBD fusion proteins and evaluated their transcriptional activities using a luciferase reporter assay. The chimeric protein had reduced repression activity compared to wild-type *BCORL1* (**Fig. 2c**). For the *CTNND1-STX5* fusion gene, the combination of non-coding exon 1 of *CTNND1* and exons 3–11 of *STX5* resulted in the deletion of 96 amino acids at the terminal end of *STX5* and increased (>twofold) *STX5* gene expression in the tumor,

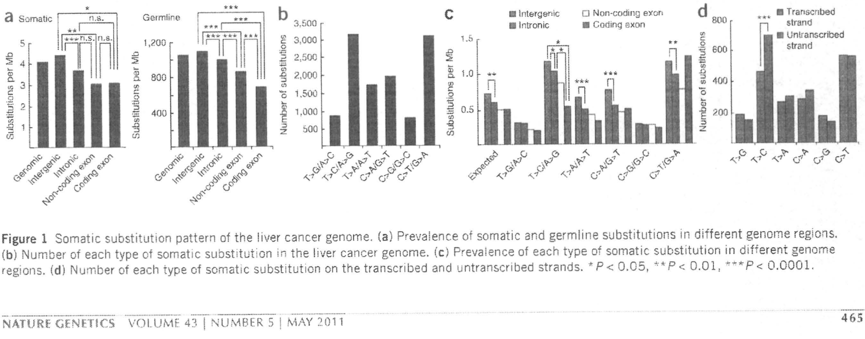

Figure 1 Somatic substitution pattern of the liver cancer genome. (**a**) Prevalence of somatic and germline substitutions in different genome regions. (**b**) Number of each type of somatic substitution in the liver cancer genome. (**c**) Prevalence of each type of somatic substitution in different genome regions. (**d**) Number of each type of somatic substitution on the transcribed and untranscribed strands. $*P < 0.05$, $**P < 0.01$, $***P < 0.0001$.

Table 2 Validated somatic non-synonymous substitutions and small indels in coding regions of a liver cancer genome

Gene	Chr.	Strand	Position	Allele change	Amino acid change	Copy number	Mutant allele (%) in whole-genome sequencing	Mutant allele (%) in whole-exome sequencing	Expression ratio (T/N)	Functional
PLEKHG5	1	–	6,452,224	G>T	Asp>Tyr	N	49.0	27.7	1.86	Deleterious
KIAA1026	1	+	15,294,007	C>A	Ala>Glu	N	45.7	nd	0.15	Tolerated
MYCL1	1	–	40,139,080	T>G	Phe>Cys	N	54.5	nd	1.93	Tolerated
PDE4B	1	+	66,231,185	C>A	Ala>Glu	N	57.1	42.9	0.83	Tolerated
CLCC1	1	–	109,284,236	A>G	Tyr>Cys	N	33.3	39.3	1.61	Deleterious
CNRIP1	2	–	68,397,833	C>T	Thr>Met	N	40.0	33.3	1.39	Deleterious
ANKRD36	2	+	97,181,397	A>G	Lys>Glu	N	17.8	nd	9.49	Tolerated
UBR3	2	+	170,511,073	A>C	Glu>Asp	N	57.1	nd	18.10	Tolerated
CUL3	2	–	225,070,790	G>A	Ser>Asn	N	42.9	52.8	12.80	Tolerated
COPS7B	2	+	232,369,129	A>G	Ile>Val	N	44.4	41.5	1.82	Tolerated
RAF1	3	–	12,625,811	A>G	Asn>Ser	N	40.0	50.0	2.31	Tolerated
ITIH3	3	+	52,813,002	A>G	Met>Val	N	43.9	nd	1.25	Deleterious
ERC2	3	–	56,148,636	G>C	Glu>Gln	N	40.0	nd	1.33	Tolerated
TBC1D23	3	+	101,496,868	del AAG	Deletion (E)	N	14.8	nd	4.90	na
ATR	3	–	143,671,657	del AT	Deletion (frame shift)	N	20.0	nd	4.49	na
SLC7A14	3	–	171,701,666	G>A	Ser>Asn	N	52.8	46.3	2.19	Deleterious
PCDH7	4	+	30,333,134	G>A	Arg>His	N	47.1	47.8	1.74	Tolerated
FAM13A	4	–	89,872,188	A>T	His>Leu	N	52.0	47.4	0.85	Tolerated
MFSD8	4	–	129,090,435	A>T	Met>Leu	Loss	62.5	74.3	1.15	Tolerated
DMGDH	5	–	78,375,996	T>A	Leu>Gln	N	50.0	37.6	3.04	Tolerated
PCDHA13	5	+	140,244,063	C>T	Pro>Ser	N	45.1	34.8	na	Deleterious
CCDC99	5	+	168,960,950	T>G	Ser>Arg	N	37.1	39.4	13.30	Deleterious
GABBR1	6	–	29,706,345	C>T	Thr>Met	N	42.0	37.8	0.59	Tolerated
CSNK2B	6	+	31,745,659	A>T	Ser>Cys	N	37.3	nd	1.41	Deleterious
MOCS1	6	–	40,003,210	G>T	Ser>Ile	N	34.4	nd	1.54	Tolerated
GTPBP2	6	–	43,699,685	A>T	Glu>Val	N	58.0	56.3	1.36	Tolerated
KHDRBS2	6	–	62,662,692	G>T	Arg>Leu	N	34.1	nd	0.88	Deleterious
SLC29A4	7	+	5,303,324	A>T	His>Leu	N	43.8	nd	7.00	Deleterious
TMEM195	7	–	15,567,887	C>G	Pro>Ala	N	41.2	38.3	1.03	Deleterious
RFC2	7	–	73,302,032	A>T	Glu>Asp	N	26.0	41.9	1.09	Tolerated
ADAM22	7	+	87,653,951	A>T	Arg>Trp	N	41.2	39.1	0.55	Deleterious
TRRAP	7	+	98,417,359	G>T	Trp>Leu	N	39.0	nd	2.07	Deleterious
XRCC2	7	–	151,977,231	G>A	Arg>Gln	N	56.2	36.5	4.18	Deleterious
MTDH	8	+	98,781,211	G>T	Val>Phe	N	33.3	46.9	14.40	Tolerated
SLA	8	–	134,141,539	C>A	Pro>Thr	N	43.6	nd	1.18	Deleterious
JAK2	9	+	5,045,703	T>G	Ile>Ser	Loss	100.0	84.2	4.84	Tolerated
NTRK2	9	+	86,532,391	G>A	Ala>Thr	Loss	90.0	85.9	0.84	Tolerated
TSC1	9	–	134,767,848	C>T	Arg>stop	Loss	13.3	13.0	1.85	Deleterious
CREM	10	+	35,496,706	A>G	Glu>Gly	N	44.8	42.3	3.28	Tolerated
C10orf95	10	–	104,200,839	T>C	Cys>Arg	N	39.7	nd	3.05	Tolerated
PSTK	10	+	124,730,061	C>T	Leu>Phe	N	53.6	nd	6.94	Deleterious
ATHL1	11	+	283,903	C>T	Ala>Val	N	40.9	26.8	1.12	Tolerated
MUC5B	11	+	1,213,214	G>T	Val>Leu	N	33.8	nd	0.83	Tolerated
DENND5A	11	–	9,181,879	C>T	Pro>Ser	N	21.4	29.9	2.43	Deleterious
GIF	11	–	59,369,438	C>T	Thr>Ile	AMP (3)	29.2	nd	0.83	Tolerated
STIP1	11	+	63,719,763	G>A	Glu>Lys	Loss	66.7	nd	1.28	Tolerated
FAT3	11	+	91,727,805	C>G	Thr>Ser	Loss	73.1	nd	na	Tolerated
PTMS	12	+	6,749,421	A>G	Glu>Gly	Loss	55.0	nd	0.56	Tolerated
ARID2	12	+	44,530,716	ins T	Insertion (frame shift)	N	31.9	nd	2.35	na
C12orf51	12	–	111,134,825	del CCTGCCACGTCA	Deletion (GDVA)	N	21.6	nd	1.44	Torelated
RBM19	12	–	112,868,641	C>T	Pro>Leu	N	49.3	42.2	1.32	Deleterious
AACS	12	+	124,142,015	G>T	Gly>Val	N	34.9	26.0	1.75	Deleterious
KHNYN	14	+	23,971,333	del CCT	Deletion (L)	N	24.1	nd	2.17	Tolerated
NOVA1	14	–	25,987,233	A>T	Leu>Phe	N	36.7	38.1	0.91	Tolerated
LTBP2	14	–	74,045,780	G>A	Gly>Glu	N	38.1	nd	3.43	Deleterious
CYFIP1	15	+	20,498,517	C>T	Ala>Val	N	55.1	41.4	1.88	Deleterious
GABRB3	15	–	24,357,328	G>T	Met>Ile	N	39.4	43.4	0.15	Tolerated
EID1	15	+	46,957,688	C>G	Ser>Cys	N	40.4	nd	8.60	Deleterious
HCN4	15	–	71,402,254	G>A	Arg>His	N	43.6	nd	0.61	Tolerated
AKAP13	15	+	84,060,152	del T	Deletion (frame shift)	N	34.5	nd	0.88	na
AXIN1	16	–	287,910	C>T	Arg>stop	Loss	78.7	nd	0.94	Deleterious
LITAF	16	–	11,554,943	del G	Deletion (frame shift)	Loss	61.3	nd	0.97	na
TP53	17	–	7,518,985	G>T	Val>Leu	Loss	78.0	73.1	0.06	Deleterious
NEK8	17	+	24,092,271	G>A	Gly>Asp	N	36.7	39.1	1.44	Deleterious
CPD	17	+	25,773,820	A>G	Tyr>Cys	N	47.1	52.3	2.28	Deleterious
LRRC30	18	+	7,221,594	C>G	Ser>Cys	N	52.0	45.6	na	Deleterious
ZNF560	19	–	9,439,794	A>C	Ile>Leu	N	58.8	48.3	0.86	Tolerated
SCRT2	20	–	593,073	T>A	Tyr>Asn	N	53.7	nd	0.51	Deleterious
USP25	21	+	16,119,227	C>T	Thr>Met	N	44.4	nd	13.00	Deleterious
USP25	21	+	16,125,626	A>C	Glu>Asp	N	35.3	38.1	na	Tolerated
ARVCF	22	–	18,341,717	C>G	Ser>Cys	N	53.0	50.0	1.30	Deleterious
USP26	X	–	131,988,824	T>C	Leu>Pro	AMP (4)	93.8	94.4	0.85	Tolerated

Except for *ANKRD36* and *TSC1*, all 63 somatic non-synonymous substitutions were predicted by whole-genome sequencing and in-house informatics method using stringent analysis criteria (Online Methods). One somatic missense substitution in *ANKRD36* was predicted under less stringent criteria. One somatic nonsense substitution in *TSC1* was predicted only by whole-exome sequencing. Chr., chromosome; N, copy neutral; AMP, amplicon; nd, not detected; na, not applicable.

Table 3 Validated somatic structural alterations in a liver cancer genome

Type	Chr. A	Break point A	CNV (Chr. A)	Chr. B	Break point B	CNV (Chr. B)	Intervening sequence	Associated genes	Fusion genes
Deletion	3	111,866,468	BCNC	3	111,868,894	BCNC	0		
Deletion	4	57,529,004	BCNC	4	57,530,452	BCNC	0	*C4orf14* (exon 4 is deleted)	
Deletion	4	92,895,135	BCNC	4	93,151,201	BCNC	0		
Deletion	5	18,130,563	BCNC	5	18,133,946	BCNC	(+) 29bp		
Deletion	6	90,130,109	BCNC	6	90,819,100	BCNC	0	*LYRM2, ANKRD6, BACH2, MDN1, CASP8AP2, RRAGD, GJA10*	
Deletion	7	69,321,043	N	7	69,404,639	N	0	*AUTS2*	
Deletion	9	132,763,157	BCNC	9	132,764,920	BCNC	0		
Deletion	10	75,477,784	BCNC	10	75,956,310	BCNC	(+) 1 bp	*AP3M1, VCL, ADK*	*VCL, ADK*
Deletion	11	67,126,436	BCNC	11	68,254,241	BCNC	0	*SUV420H1, SAPS3, ACY3, ALDH3B2, CHKA, TCIRG1, LRP5, GAL, ALDH3B1, TBX10, NDUFV1, UNC93B1, NUDT8, C11orf24*	
Deletion	15	47,394,203	BCNC	15	47,467,920	BCNC	0	*GALK2, C15orf33*	
Deletion	17	15,902,440	BCNC	17	16,056,159	BCNC	0	*NCOR1* (homozygous deletion)	
Inversion	4	60,946,299	N	4	60,947,151	N	0		
Inversion	4	172,703,199	Loss	4	172,706,239	Loss	(+) 4bp		
Inversion	11	57,305,269	BCNC	11	62,352,275	BCNC	0	*CTNND1* (UTR), *STX5*	*CTNND1, STX5*
Inversion	11	57,770,822	BCNC	11	67,133,985	BCNC	0	*NDUFV1*	
Inversion	11	62,309,952	BCNC	11	70,746,006	BCNC	0	*TAF6L*	
Inversion	11	69,067,231	AMP	11	69,317,424	AMP	0		
Inversion	11	69,093,978	AMP	11	69,098,117	AMP	0		
Inversion	11	69,871,206	AMP	11	69,877,391	AMP	(+) 6bp	*PPFIA1*	
Inversion	X	129,015,072	N	X	129,029,501	BCNC	(+) 23bp	*BCORL1, ELF4*	*BCORL1, ELF4*
Inversion	X	129,016,981	N	X	129,031,425	BCNC	0	*BCORL1, ELF4*	*BCORL1, ELF4*
Tandem duplication	11	67,043,308	BCNC	11	67,318,685	BCNC	0	*ACY3, ALDH3B2, GSTP1, TBX10, NDUFV1, NUDT8, CABP2, LOC645332*	*CABP2, LOC645332*
Translocation	11	69,316,960	AMP	X	129,030,346	BCNC	0	*ELF4*	

The inversions at Xq25 occurred from one rearrangement event and the total number of inversion is counted as nine. Chr., chromosome; BCNC, boundary of copy number change; N, copy neutral; AMP, amplicon.

which harbors only the rearranged allele (**Fig. 2d** and **Supplementary Fig. 9**). We screened for the presence of these four chimera transcripts by RT-PCR, but we detected no recurrent fusion event in 47 cases of primary HCC, possibly due to the low frequency of these rearrangements in HCC or because of the technical difficulty in detecting all variant fusion transcripts.

We also sequenced the whole exomes of the same samples using an in-solution gene enrichment system[5] (**Fig. 3a**). Capture probes for whole-exome sequencing were designed to cover the protein coding exons using the consensus coding sequences, excluding highly homologous regions. The average coverage of the whole exome sequences (41.3 Mb in total) was about twice (76.8× for HCC and 74.3× for lymphocytes) that of the whole genome sequences and had one twelfth of the total sequence amount (8.9 Gb for HCC and 8.6 Gb for lymphocyte) (**Supplementary Table 3**). Whole-exome sequencing detected 47 non-synonymous somatic substitutions, 40 of which were validated by Sanger sequencing. Among the validated substitutions, a nonsense substitution (p.Arg785X) in *TSC1*, located in the hemizygous region (9q34), was not detected by whole-genome sequencing (**Fig. 3b**). Capillary sequencing validated the same substitution with a very low

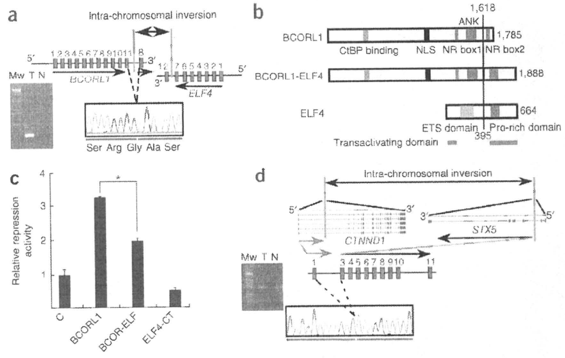

Figure 2 Characterization of rearrangements in liver cancer. (**a**) Top, schematic representation of the intra-chromosomal inversion at Xq25. Bottom left, RT-PCR analysis of the fused *BCORL1-ELF4* transcript in tumor (T) and non-cancerous liver (N) tissues. We detected no *ELF4-BCORL1* transcript (data not shown). Bottom right, sequence chromatography of the fusion transcript revealed an in-frame protein. Mw, molecular marker. (**b**) Schematic representation of the BCORL1-ELF4 fusion protein. BCORL1 (top) contains a CtBP1 binding domain (PXDLS sequence), a binuclear localization signal (NLS), two LXXLL nuclear receptor recruitment motifs (NR box) and tandem ankyrin repeats (ANK). ELF4 (bottom) contains an ETS (E Twenty Six) DNA binding domain and a proline-rich domain. Transactivating domains are indicated by the red bars[16]. The BCORL1-ELF4 chimeric protein includes most of BCORL1 (1–1,618 amino acids) lacking the NR box2 and the carboxyl-terminal portion of ELF4 containing the proline-rich domain. The number of amino acids is indicated on the right. (**c**) Wild-type BCORL1, ELF4-CT (395–664 amino acids) and the BCORL1-ELF4 chimera were expressed as Gal4-DBD fusion proteins, and their relative transcriptional activities were compared to the Gal4-DBD protein (C) as shown. (**d**) Characterization of the *CTNND1-STX5* fusion gene. Bottom left, RTPCR analysis of the fused *CTNND1-STX5* transcript in tumor (T) and non-cancerous liver tissue (N). Bottom right, sequence chromatography of the fusion transcript. Data is the mean ± s.d. ($n = 3$). $*P < 0.001$.

5. Gnirke, A. *et al.* Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. *Nat. Biotechnol.* **27**, 182–189 (2009).
6. Pleasance, E.D. *et al.* A comprehensive catalogue of somatic mutations from a human cancer genome. *Nature* **463**, 191–196 (2010).
7. Xiang, Z. *et al.* Identification of somatic *JAK1* mutations in patients with acute myeloid leukemia. *Blood* **111**, 4809–4812 (2008).
8. Pagan, J.K. *et al.* A novel corepressor, BCoR-L1, represses transcription through an interaction with CtBP. *J. Biol. Chem.* **282**, 15248–15257 (2002).
9. Miyazaki, Y., Sun, X., Uchida, H., Zhang, J. & Nimer, S. MEF, a novel transcription factor with an Elf-1 like DNA binding domain but distinct transcriptional activating properties. *Oncogene* **13**, 1721–1729 (1996).
10. Suico, M.A. *et al.* Functional dissection of the ETS transcription factor MEF. *Biochim. Biophys. Acta* **1577**, 113–120 (2002).
11. Mardis, E.R. *et al.* Recurring mutations found by sequencing an acute myeloid leukemia genome. *N. Engl. J. Med.* **361**, 1058–1066 (2009).
12. Ding, L. *et al.* Genome remodelling in a basal-like breast cancer metastasis and xenograft. *Nature* **464**, 999–1005 (2010).
13. Shah, S.P. *et al.* Mutation evolution in a lobular breast tumour profiled at single nucleotide resolution. *Nature* **461**, 809–813 (2009).
14. Parsons, D.W. *et al.* An integrated genomic analysis of human glioblastoma multiforme. *Science* **321**, 1807–1812 (2008).
15. Jones, S. *et al.* Core signaling pathways in human pancreatic cancers revealed by global genomic analyses. *Science* **321**, 1801–1806 (2008).
16. Wood, L.D. *et al.* The genomic landscapes of human breast and colorectal cancers. *Science* **318**, 1108–1113 (2007).
17. Pleasance, E.D. *et al.* A small-cell lung cancer genome with complex signatures of tobacco exposure. *Nature* **463**, 184–190 (2010).
18. Lee, W. *et al.* The mutation spectrum revealed by paired genome sequences from a lung cancer patient. *Nature* **465**, 473–477 (2010).
19. Greenman, C. *et al.* Patterns of somatic mutation in human cancer genomes. *Nature* **446**, 153–158 (2007).
20. Machida, K. *et al.* Hepatitis C virus induces a mutator phenotype: enhanced mutations of immunoglobulin and protooncogenes. *Proc. Natl. Acad. Sci. USA* **101**, 4262–4267 (2004).
21. Kim, M.Y. *et al.* Tumor self-seeding by circulating cancer cells. *Cell* **139**, 1315–1326 (2009).
22. Guertin, D.A. & Sabatini, D.M. Defining the role of mTOR in cancer. *Cancer Cell* **12**, 9–22 (2007).
23. Yilmaz, O.H. *et al.* Pten dependence distinguishes haematopoietic stem cells from leukaemia-initiating cells. *Nature* **441**, 475–482 (2006).
24. Meric-Bernstam, F. & Gonzalez-Angulo, A.M. Targeting the mTOR signaling network for cancer therapy. *J. Clin. Oncol.* **27**, 2278–2287 (2009).

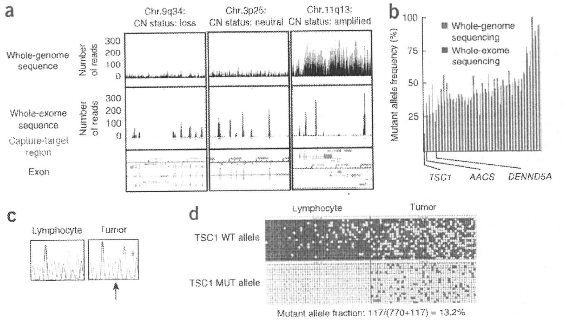

Figure 3 Intra-tumoral genetic heterogeneity detected by exon-capture sequencing. (**a**) Specific enrichment and high sequence coverage of the target genome regions indicated by the sequence viewer (copy number (CN) status is shown above). The distribution and number of reads (black, forward read; gray, reverse read) from whole-genome sequencing (top) and whole-exome sequencing (middle) are shown. The location of the capture target regions (red box) and the exons (green box) along the genome are shown at the bottom. Note that the number of reads is dependent on copy number status. (**b**) Mutant allele frequency detected by whole-genome sequencing and whole-exome sequencing. *TSC1*, *AACS* (whose heterogeneity is shown in **Supplementary Fig. 10**) and *DENND5A* are indicated. (**c**) *TSC1* mutation in the liver cancer subpopulation. Sequence chromatograms of *TSC1* in lymphocytes and whole-tumor tissue are shown. Note the small peak for the mutant T allele (indicated by the arrow) in the tumor DNA. (**d**) Determination of mutant *TSC1* allele frequency by digital PCR genotyping. WT, wild type; MUT, mutant.

signal peak (**Fig. 3c**), and digital genotyping showed that 13.2% of the tumor alleles harbored this substitution (**Fig. 3d**), suggesting that this substitution occurred in a minor population of cancer cells. Whole-exome sequencing missed 25 non-synonymous somatic substitutions that were detected by whole-genome sequencing. These missed substitutions were located in regions where sequence coverage was low or where further optimization of the probe design was required.

The number of non-synonymous somatic substitutions validated in this HCC (63) was greater than those for acute myeloid leukemia[11] (10), basal-like breast cancer[12] (22), lobular carcinoma[13] (32), glioblastoma multiforme[14] (32) and pancreatic cancer[15] (43) but is in the range of those previously reported for colorectal[16] (70) and breast[16] (88) cancer. We have shown that the pattern of somatic substitutions in a HCV-associated HCC genome is different (predominance of T>C, especially at ApT sites, and C>T, especially at CpG sites) compared to smoking-related[17,18] and ultraviolet light–related[6] cancers. Preferential C>T/G>A transition may partly be due to the higher frequency of CpG methylation in the genome sequence and is a common form of mutation in cancers[19]. Therefore, the T>C/A>G transition could be a characteristic mutational signature of HCV-associated cancer, which would be consistent with a previous observation that HCV induces error-prone DNA polymerases that preferentially cause the T>C/A>G mutation[20]. It is also possible that this mutation pattern is independent of viral infection and is organ specific, as a comparable substitution spectrum has been reported in renal cancer[19]. Additionally, only T>C changes, but not C>T changes, were effectively repaired on the transcribed strand. Similar enhanced transcription-coupled repair on preferentially acquired substitutions has been reported in other cancers[6,17,18] and could be a common phenomenon in cancer mutation.

Because single-molecule sequencing has the capability to detect every individual somatic event in parallel, higher sequence coverage will enable us to clarify the intra-tumoral heterogeneity that is associated with diverse aspects of clinical behavior such as metastasis[21]. The *TSC1* complex, which is inactivated in a subpopulation of tumors, negatively regulates the mammalian target of rapamycin signaling, which is an important oncogenic pathway related to the growth, metabolism and stemness of cancer cells[22,23], and could be a promising molecular therapeutic target in HCC progression[24].

URLs. International Cancer Genome Consortium, http://www.icgc.org/; Catalogue of Somatic Mutations in Cancer, http://www.sanger.ac.uk/genetics/CGP/cosmic/; BLASTN, ftp://ftp.ncbi.nlm.nih.gov/blast/executables/release/LATEST.

METHODS

Methods and any associated references are available in the online version of the paper at http://www.nature.com/naturegenetics/.

Note: Supplementary information is available on the Nature Genetics website.

ACKNOWLEDGMENTS

We thank K.K. Khanna (The Queensland Institute of Medical Research) for providing a human *BCORL1* cDNA clone; T.D. Taylor (RIKEN) for comments on the manuscript; T. Urushidate, S. Ohashi, S. Ohnami, A. Kokubu, N. Okada, K. Shiina, H. Meguro and K. Nakano for their excellent technical assistance. This work was supported by the Program for Promotion of Fundamental Studies in Health Sciences of the National Institute of Biomedical Innovation (NIBIO), Japan, and the Industrial Technology Research Grant Program from the New Energy and Industrial Technology Development Organization (NEDO), Japan. This study is associated with the International Cancer Genome Consortium (ICGC), and the mutation data were deposited at and released from the ICGC web site.

AUTHOR CONTRIBUTIONS

The study was designed by T. Shibata, H.A., T.Y. and J.K. Sequencing and data analyses were conducted by Y.T., K.T., S.Y., S.T., K. Sonoda and H.T. Allele typing and copy number analyses were performed by H.S. and S.I. Other molecular studies were done by Y.A., F.H., T. Shirakihara, and L.W.; H.O., K. Shimada, T.K., T.O. and K.K. coordinated collection of clinical sample and information. The manuscript was written by Y.T., T. Shibata, K.T., S.Y., H.A. and T.Y.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Published online at http://www.nature.com/naturegenetics/.
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions/.

1. El-Serag, H.B. & Rudolph, K.L. Hepatocellular carcinoma: epidemiology and molecular carcinogenesis. *Gastroenterology* **132**, 2557–2576 (2007).
2. Bentley, D.R. *et al.* Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**, 53–59 (2008).
3. Huang, W., Sherman, B.T. & Lempicki, R.A. Systematic and integrative analysis of large gene lists using DAVID Bioinformatics Resources. *Nat. Protoc.* **4**, 44–57 (2009).
4. Ng, S.B. *et al.* Targeted capture and massively parallel sequencing of 12 human exomes. *Nature* **461**, 272–276 (2009).

ONLINE METHODS

Whole-genome sequencing. High molecular weight DNA was extracted from freshly frozen tumor tissue and lymphocytes. DNA was fragmented using an ultrasonic solubilizer (Covaris) using a combination of quick bursts (20% duty, 5 intensity with 200 cycles per burst for 5 s) and sonication (10% duty, 5 intensity with 200 cycles per burst for 120 s) for the short fragment DNA library. DNA of the appropriate size was gel purified to exclude any inappropriate DNA fusions during library construction. The short fragment DNA libraries were generated using a paired-end DNA sample prep kit (Illumina) following the manufacturer's protocols. The concentration of the libraries was quantified using a Bioanalyzer (Agilent Technologies); 4~8 pM/lane of DNA was applied to the flow cell, and paired-end sequencing was performed using the GAIIx sequencer (Illumina).

Whole-exome capture sequencing. Whole-exome capture sequencing was performed using the SureSelect Target Enrichment System (Agilent Technologies) in accordance with the manufacturer's protocol with slight modifications. Briefly, the same Illumina sequence libraries as those prepared for the whole-genome sequence were amplified with six cycles of PCR, and then 500 ng of the amplified libraries was hybridized with the capture probes for 24 h. The hybridized sequence libraries were collected and further amplified with 14 cycles of PCR. We generated 51-nucleotide–long paired-end reads using the GAIIx sequencer (Illumina). We used five lanes of a paired-end flow cell for each sample.

Bioinformatics (Supplementary Fig. 11). *Sequence alignment to the human genome and removal of PCR duplications.* Paired-end reads were aligned to the human reference genome (hg18, NCBI Build 36.1) using Burrows-Wheeler Aligner (BWA) (version 0.4.9)[25]. Because there were duplicated reads which were generated during the PCR amplification process, paired-end reads that aligned to the same genomic positions were removed using SAMtools (version 0.1.5c)[26] and a program developed in house. We removed 12.5% (14.6/117.1 Gbp) of the aligned reads for tumor and 7.1% (6.1/86.3 Gbp) for lymphocytes.

Detection of somatic single nucleotide variations (SNVs) **(Supplementary Fig. 12).** Based on the genotyping data from two SNP arrays, appropriate thresholds for base quality, mapping quality and frequency of non-reference alleles were determined to obtain the highest confidence calls for SNV detection (**Supplementary Table 4**). To predict somatic SNVs, the alignment results were classified, and three datasets were constructed. Dataset 1 included paired-end reads with both ends aligned uniquely and with proper spacing and orientation. Dataset 2 included paired-end reads that aligned uniquely for at least one read and with proper spacing and orientation of the reads. Dataset 3 included dataset 2 and paired-end reads for which both ends aligned uniquely but with improper spacing or orientation or both. Dataset 1 likely contains false positive somatic SNVs because of the low sequence depth of the lymphocyte genome, and dataset 3 likely contains false positives due to misalignments of the sequence reads. To reduce the number of false positives, the following filters were applied to these three datasets, and concordant somatic SNVs among the three datasets were selected: (i) a mapping quality score of 20 was used as a cutoff value for read selection; (ii) base quality scores of 10 and 15 were used as cutoff values for base selection for the tumor and lymphocyte genomes, respectively; (iii) SNVs were selected when the frequency of the non-reference allele was at least 15% in the tumor genome and 5% in the lymphocyte genome; (iv) SNVs located within 5 bp from a potential insertion or deletion were discarded; (v) SNVs with a root mean square mapping quality score of the reads covering the SNV less than 40 were discarded; (vi) when there were three or more SNVs within any 10-bp window, all of them were discarded; (vii) SNVs with a consensus quality score less than 20 as calculated by SAMtools (version 0.1.5c) were discarded; (viii) when a base with a consensus quality score less than 20 was located within 3-bp on either side of a SNV, the SNV was discarded; (ix) for the tumor genome, SNVs found in at least two sequence reads with the same SNV were selected; (x) for the lymphocyte genome, SNVs covered by at least six sequence reads were selected; and (xi) the repetitive regions within 1 Mb of a centromeric or telomeric sequence gap were excluded. By comparing the predicted nucleotide variations in the tumor and lymphocyte genomes, somatic SNVs which occurred only in the tumor genome were identified. If somatic SNVs were not covered in the lymphocyte genome by at least six sequence reads, they were discarded.

Using this approach, 66 non-synonymous and 24 synonymous somatic SNVs in protein-coding regions were predicted. These 90 substitutions were examined by Sanger sequencing of both the tumor and lymphocyte genomes, and 81 of them were validated as somatic mutations. Of the remaining nine substitutions, three could not be amplified by PCR, four could not be sequenced because of the surrounding repetitive sequences, and two could not be validated likely because they were located in highly homologous segmentally duplicated or processed pseudogene regions, suggesting a high prediction accuracy (specificity, 81/83 = 97.6%) for our approach for detecting somatic SNVs in protein-coding regions. An additional 36 non-synonymous somatic SNVs were also predicted using only dataset 3 and filtering methods (i–iv) (less stringent filtering condition). Five of these SNVs were not validated and 30 of them were found to be germline variations by Sanger sequencing, and only the one remaining was validated as a somatic mutation. These findings suggest that our filtering method (stringent condition) effectively removed false-positive somatic SNVs.

Detection of somatic structural alterations. To detect structural alterations, paired-end reads for which both ends aligned uniquely to the human reference genome, but with improper spacing or orientation or both, were used. First, paired-end reads were selected based on the following filtering conditions: (i) sequence reads with mapping quality scores greater than 37; and (ii) sequence reads aligned with two mismatches or less.

Rearrangements were then identified using the following analytical conditions: (i) 'clusters' which included reads aligned within the maximum insert distance were constructed from the forward and reverse alignments, respectively (two reads were allocated to the same cluster if their end positions were not further apart than the maximum insert distance); (ii) clusters whose distance between the leftmost and rightmost reads were greater than the maximum insert distance were discarded; (iii) paired-end reads were selected if one end sequence was allocated in the 'forward cluster' and the other end was allocated in the 'reverse cluster' (we called these 'forward cluster and reverse cluster' paired clusters); (iv) if a cluster overlapped another cluster, all of the overlapping paired-clusters were discarded; (v) for the tumor genome, rearrangements (paired-clusters) predicted by at least four paired-end reads which included at least one paired-end read perfectly matched to the human reference genome were selected; and (vi) for the lymphocyte genome, rearrangements (paired clusters) predicted by at least one paired-end read were selected. By comparing the predicted rearrangements in the tumor and lymphocyte genomes, somatic rearrangements that were only detected in the tumor genome were identified.

Lastly, rearrangements predicted due to variations in the analyzed genomes were removed. For this analysis, paired-end reads contained in paired clusters were aligned to the human reference genome using the BLASTN program (see URLs). If one end sequence was aligned to the region of paired clusters (the flanking region of the rearrangement breakpoint) and the other end was aligned with proper spacing and orientation, the rearrangement was removed. An expectation value of 1,000 was used as a cutoff value for BLASTN so that paired-end reads with low similarity to the human reference genome could also be aligned.

Using this method, 33 somatic rearrangements were predicted and 22 of these were validated by Sanger sequencing of the rearrangement breakpoints in both the tumor and lymphocyte genomes.

Exome capture sequence analysis. To analyze the capture sequencing data, the Illumina sequencing pipeline version 1.4 and in-house programs were used. The sequence reads were mapped to the human reference sequence (NCBI Build 36.3) using GERALD (Illumina), and only high-quality ('pass filter') reads with base-call quality scores more than ten were used for SNV detection.

SNVs were determined using the frequency (>20%) of the highest non-reference base call with a read depth greater than 20×.

Other molecular analyses. SNP genotyping and copy number detection were determined using the Affymetrix Mapping 500K Array, the Agilent Human Genome CGH microarray and the Illumina Human 610-Quad BeadChip system. Gene expression levels of the tumor were measured using the Agilent Whole Human Genome Oligo Microarray. Wild-type and mutant allele frequencies were determined using the Digital PCR system.

Detailed experimental methods and additional bioinformatics procedures are described in **Supplementary Note**. The somatic substitutions and insertions/deletions found are listed in **Supplementary Tables 5–9**.

25. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler Transform. *Bioinformatics* **25**, 1754–1760 (2009).
26. Li, H. *et al.* The Sequence alignment/map (SAM) format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).

doi:10.1038/ng.804

ORIGINAL ARTICLE

Genome-Wide Association Study on Overall Survival of Advanced Non-small Cell Lung Cancer Patients Treated with Carboplatin and Paclitaxel

Yasunori Sato, PhD,[*][†] Noboru Yamamoto, MD,[‡] Hideo Kunitoh, MD,[‡][§] Yuichiro Ohe, MD,[‡] Hironobu Minami, MD,[||][¶] Nan M. Laird, PhD,[†] Noriko Katori, PhD,[#] Yoshiro Saito, PhD,[**] Sumiko Ohnami, BS,[*] Hiromi Sakamoto, PhD,[*] Jun-ichi Sawada, PhD,[††] Nagahiro Saijo, MD, PhD,[‡‡] Teruhiko Yoshida, MD, PhD,[*] and Tomohide Tamura, MD, PhD[‡]

Purpose: Our goal was to identify candidate polymorphisms that could influence overall survival (OS) in advanced non-small cell lung cancer (NSCLC) patients treated with carboplatin (CBDCA) and paclitaxel (PTX).

Methods: Chemotherapy-naïve stage IIIB or IV NSCLC patients treated with CBDCA (area under the curve = 6 mg/mL/min) and PTX (200 mg/m^2, 3-hour period) were eligible for this study. The DNA samples were extracted from peripheral blood mononuclear cells before treatment, and genotypes at approximately 110,000 gene-centric single-nucleotide polymorphisms (SNPs) were obtained by Illumina's Sentrix Human-1 Genotyping BeadChip. Statistical analyses were performed by the log-rank test and Cox proportional hazards model.

Results: From July 2002 to May 2004, 105 patients received a total of 308 cycles of treatment. The median survival time (MST) of 105 patients was 17.1 months. In the genome-wide association study, three SNPs were associated significantly with shortened OS after multiple comparison adjustment: rs1656402 in the *EIF4E2* gene (MST was 18.0 and 7.7 months for AG [$n = 50$] + AA [$n = 40$] and GG [$n = 15$], respectively; $p = 8.4 \times 10^{-8}$), rs1209950 in the *ETS2* gene (MST = 17.7 and 7.4 months for CC [$n = 94$] and CT [$n = 11$] + TT [$n = 0$]; $p = 2.8 \times 10^{-7}$), and rs9981861 in the *DSCAM* gene (MST = 17.1 and 3.8 months for AA [$n = 75$] + AG [$n = 26$] and GG [$n = 4$]; $p = 3.5 \times 10^{-6}$).

Conclusion: Three SNPs were identified as new prognostic biomarker candidates for advanced NSCLC treated with CBDCA and PTX. The agnostic genome-wide association study may unveil unexplored molecular pathways associated with the drug response, but our findings should be replicated by other investigators.

Key Words: Advanced non-small lung cancer, Carboplatin, Paclitaxel, Genome-wide association study, Single-nucleotide polymorphisms.

(*J Thorac Oncol.* 2011;6: 132–138)

Lung cancer is the leading cause of cancer death in Japan and worldwide for both men and women.[1] Non-small cell lung cancer (NSCLC) accounts for approximately 85% of lung cancer cases. Several third-generation agents are available for the treatment of NSCLC, including docetaxel, paclitaxel (PTX), gemcitabine, and vinorelbine, and the combination of one of these agents with a platinum compound has been considered the standard treatment option for advanced NSCLC.[2–9]

Despite these advances, survival prospects still remain disappointingly low for most patients. To seek further improvements in response rate and survival time, the conventional treatment approach to NSCLC is beginning to shift toward the application of specific strategies and techniques, such as pharmacogenomics to tailor treatment to individual patients.[10,11]

To identify the clinical predictors of outcome, it is critically important to observe individual differences in drug response and the role of genetic polymorphisms that are relevant to the pathways of drug metabolism and/or the biology of drug responses. However, genetic polymorphisms that are associated with overall survival (OS) or antitumor effect have not yet been fully elucidated.

With this as background, this prospective study employed a genome-wide association study (GWAS) to identify candidate polymorphisms that could influence OS in advanced NSCLC patients treated with carboplatin (CBDCA) and PTX. Possible associations with toxicities and pharma-

[*] Genetics Division, National Cancer Center Research Institute, Tokyo, Japan; [†] Department of Biostatistics, Harvard School of Public Health, Boston, Massachusetts; [‡] Division of Internal Medicine, National Cancer Center Hospital; [§] Department of Respiratory Medicine, Mitsui Memorial Hospital, Tokyo; [||] Division of Internal Medicine, National Cancer Center Hospital East, Chiba; [¶] Division of Oncology/Hematology, Kobe University Graduate School of Medicine, Kobe; [#] Divisions of Drugs, [**] Medicinal Safety Science, and [††] Functional Biochemistry and Genomics, National Institute of Health Sciences, Tokyo; and [‡‡] National Cancer Center Hospital East, Chiba, Japan.

Disclosure: Dr. Minami has received honororia from Bristol-Myers Squibb KK. The other authors declare no conflicts of interest.

Address for correspondence: Teruhiko Yoshida, MD, Genetics Division, National Cancer Center Research Institute, 5-1-1 Tsukiji, Chuo-ku, Tokyo 104-0045, Japan. E-mail: tyoshida@ncc.go.jp

The first two authors contributed equally to this work.

ISSN: 1556-0864/11/0601-0132

cokinetic (PK) parameters were also tested to complement our previous candidate gene approach focusing on CYP3A4[12] and CYP2C8.[13]

PATIENTS AND METHODS

Patient Recruitment and Treatment Schedule

Patients with histologically and/or cytologically documented NSCLC were eligible for participation in the study and treated with CBDCA and PTX at the National Cancer Center Hospital and National Cancer Center Hospital East. Each patient had to meet the following criteria: clinical stage IIIB or IV, no prior chemotherapy, no prior surgery and/or radiotherapy for the primary site, age older than 20 years, and Eastern Cooperative Oncology Group performance status[14] between 0 and 2. This study was approved by the Ethics Review Committees of the National Cancer Center and National Institutes of Health Sciences, and written informed consent was obtained from all patients before study entry.

One hundred five patients received 200 mg/m^2 of PTX (Bristol-Myers K.K., Tokyo, Japan) over a 3-hour period followed by carboplatin at a dose calculated to produce an area under the concentration time curve of 6.0 mg/mL/min on day 1, with the cycle being repeated every 3 weeks. In addition, to prevent hypersensitivity reactions, all patients received short-term premedication including dexamethasone, ranitidine, and an antiallergic agent (diphenhydramine or chlorpheniramine maleate).

Monitoring, Response and Toxicity Evaluation, and Follow-Up

A complete medical history and data on physical examinations were recorded before the CBDCA and PTX combination therapy. Complete blood cell and platelet counts as well as blood chemistry were measured once a week during the first 2 months of the treatment. Response was evaluated according to the Response Evaluation Criteria in Solid Tumors (RECIST), except that tumor markers were excluded from the criteria. Toxicity grading criteria in National Cancer Institute Common Toxicity Criteria Version 2.0 were used to evaluate toxicity. Patients were followed by direct evaluation or resident registration until death or up to 5 years after treatment. OS was calculated from the date of patient enrollment in this study to the date of death or the last follow-up.

Pharmacokinetic Sampling and Analysis

For PTX PK analysis, 5 ml of heparinized blood was sampled before the first PTX administration and at 0, 1, 3, and 9 hours after the termination of the infusion. The area under the curve (AUC) and clearance (CL m^{-2}) were calculated by a curve fitting method using the model of two compartments with constant infusion using WinNonlin ver. 3.3 (Pharsight Corporation, Mountain View, CA). The PK data were used in our previous pharmacogenetic analyses.[12,13]

DNA Extraction and Genotyping

Whole blood was collected from patients at the time of enrollment, and DNA was extracted from peripheral lymphocytes using a proteinase-K phenol chloroform method or Qiagen FlexiGene DNA isolation kit (QIAGEN Inc., Valencia, CA). All samples were assayed with the Illumina Infinium Human-1 BeadChip (Illumina Inc., San Diego, CA), which assays 109,365 gene-centric single-nucleotide polymorphisms (SNPs). If a genotyping call rate on all SNPs was found to be less than 95%, the sample was excluded from the analysis.

Statistical Analysis

As a quality control for genotyping, Hardy-Weinberg equilibrium testing was applied. To estimate the association between OS and genotypes, hazard ratios (HRs) and 95% confidence intervals were calculated using univariate or multivariate Cox proportional hazards models[15,16] and assessed using the log-rank test. Survival curves were drawn using the Kaplan-Meier method.[14] Statistical significance level was set to 0.05, two sided, after Holm's adjustment for a multiple testing.[17] All statistical analyses were performed with the use of SAS software, version 9.1.3 (SAS Institute Inc., Cary, NC). All statistical analyses were planned before the study.

RESULTS

Patient Characteristics, Survival, Response, and Toxicity

From July 2002 to May 2004, 239 patients treated with PTX were enrolled. Among them, 110 chemotherapy-naïve advanced NSCLC patients treated with CBDCA (AUC = 6 mg/mL/min) and PTX (200 mg/m^2, 3-hour period) were eligible in this study, but five patients were excluded from the analysis because genotyping data were not available. Their characteristics are shown in Table 1. All patients were followed up for more than 2.5 years, and the median follow-up time among censored observations was 38 months (range, 27–46 months), with 89 patients deceased (85%) as of November 2006. The median survival time (MST) of the 105 patients was 17.1 months (95% confidence interval: 15.0–18.7) (Figure 1). The 1- and 3-year survival probabilities were 68% and 16%, respectively.

Of the 105 patients, changes in tumor measurements were partial response in 43 (41%) patients, stable disease in 47 (45%), progressive disease in 11 (10%), and not evaluated in 4 (4%). There were no cases with a complete response.

All patients were evaluated for toxicity. Hematologic toxicity and nonhematologic toxicity are summarized in Table 2. Grade 3 or 4 nonhematologic toxicity occurred in 15

TABLE 1. Patient Characteristics

Assessable patients	105
Gender (male/female)	76/29
Age, median (range)	61 (29–80)
PS (0/1/2)	20/82/3
Stage (IIIB/IV)	46/59
No. of treatment cycles	
Mean	2.93
Range	1.0–6.0

PS, performance status.

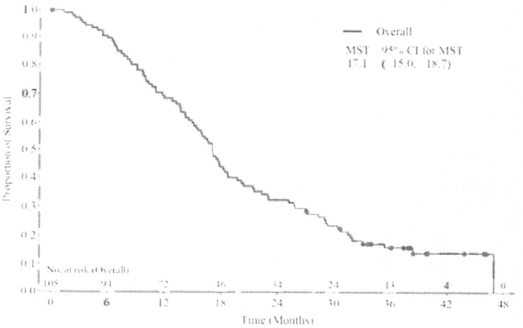

FIGURE 1. Kaplan-Meier plot for overall survival.

TABLE 2. Incidence of Hematologic and Nonhematologic Toxicities After the First Cycle

Toxicity	Grade 1	Grade 2	Grade 3	Grade 4	Total
Leukopenia	40	34	9	0	101
Neutropenia	8	22	39	18	105
Anemia	73	16	2	0	105
Thrombocytopenia	16	3	0	0	102
Febrile neutropenia	0	0	5	0	105
Nausea	7	3	0	0	105
Vomiting	8	4	3	0	105
Diarrhea	5	6	0	1	105
Arthralgia	58	12	2	0	105
Myalgia	47	10	1	0	105
Hyperbilirubinemia	33	10	0	0	105
AST (GOT) increase	38	1	0	0	105
ALT (GPT) increase	38	3	1	0	105
ALP increase	32	5	0	0	105
Neuropathy, sensory	65	6	1	0	105
Neuropathy, motor	1	0	0	1	105

AST, aspartate transaminase; GOT, glutamic oxaloacetic transaminase; ALT, alanine aminotransferase; GPT, glutamate pyruvate transaminase; ALP, alkaline phosphatase.

(14%) patients, suggesting that nonhematologic toxicity was generally mild; but grade 4 motor neuropathy occurred in one patient and grade 4 diarrhea occurred in another. On the other hand, grade 3 or 4 hematologic toxicity occurred in 57 (53%) patients. Grade 4 neutropenia occurred in 18 (17%) patients. Febrile neutropenia (grade 3) occurred in five patients.

Effects of Patients' Background on Overall Survival

The effects of patients' background on OS were analyzed as summarized in Table 3. The effects of gender, Eastern Cooperative Oncology Group performance status, and tumor response showed significant associations with OS, but age, stage, and number of cycles did not show a significant association.

TABLE 3. Univariate Analysis of Patients' Characteristics

	Overall Survival		
Variable	Crude HR	95% CI for HR	p
Age			
≥65 vs. <65	1.12	0.72–1.71	0.61
Gender			
Male vs. female	2.06	1.26–3.39	0.0039
PS			
2 vs. 0–1	7.68	2.28–25.8	0.0010
Stage			
IV vs. IIIB	1.19	0.78–1.83	0.40
No. of cycles	0.92	0.74–1.13	0.42
Tumor response			
PR vs. PD	0.199	0.098–0.403	<.0001
NC vs. PD	0.216	0.108–0.434	<.0001

CI, confidence interval; HR, hazard ratio; PR, partial response; PD, progressive disease; NC, no change.

Pharmacogenomic Analyses

Table 4 lists 10 SNPs, showing the least p values for log-rank test. The following three SNPs were associated significantly with shortened OS after multiple comparison adjustment: rs1656402 in the *EIF4E2* gene (MST for AG [n = 50] + AA [n = 40] and GG [n = 15] were 18.0 and 7.7 months, respectively; $p = 8.4 \times 10^{-8}$, HR = 4.22 [2.32–7.66]), rs1209950 in the *ETS2* gene (MST for CC [n = 94] and CT [n = 11] + TT [n = 0] were 17.7 and 7.4 months, respectively; $p = 2.8 \times 10^{-7}$, HR = 4.96 [2.52–9.76]), and rs9981861 in the *DSCAM* gene (MST for GG [n = 75] + AG [n = 26] and AA [n = 4] were 17.1 and 3.8 months, respectively; $p = 3.5 \times 10^{-6}$, HR = 16.1 [5.38–51.2]). In Figure 2, the Kaplan-Meier plots were drawn with subjects stratified into subgroups according to each significant polymorphism in either dominant or recessive model. Two (rs1656402 and rs9981861) of these significant SNPs were associated with tumor response and AUC 6α-,C3′-p-dihydroxy-PTX as shown

TABLE 4. Ten SNPs Associated with OS in GWAS

		SNP Information			Patients						
Chr #	Rs #	Gene Symbol	Genotype	Frequency	Total	Events	MST (95% CI)	HR (95% CI)	p^a	p^b	p^c
2	rs1656402	*EIF4E2*	AA	0.145	40	37	15.6 (13.5–17.0)	Ref	8.4×10^{-8}	4.5×10^{-7}	0.0046
			AG	0.461	50	37	24.4 (18.6–30.3)	0.42 (0.26–0.67)			
			GG	0.393	15	15	7.69 (5.95–12.7)	2.73 (1.46–5.10)			
21	rs1209950	*ETS2*	CC	0.938	94	78	17.6 (16.2–21.4)	Ref	2.8×10^{-7}	6.5×10^{-5}	0.015
			CT	0.059	11	11	7.39 (4.86–10.2)	4.96 (2.52–9.76)			
			TT	0.002	—	—	—	NA			
21	rs9981861	*DSCAM*	AA	0.652	75	61	17.8 (15.3–21.4)	Ref	3.5×10^{-6}	9.2×10^{-7}	0.050
			AG	0.314	26	24	16.5 (2.14–18.1)	1.33 (0.82–2.15)			
			GG	0.034	4	4	3.78 (2.14–7.69)	18.0 (5.78–56.2)			
2	rs10496036	*RTN4*	GG	0.701	84	70	17.6 (15.9–21.4)	Ref	2.4×10^{-5}	0.00063	1.00
			AG	0.270	18	2	14.1 (9.63–19.6)	1.52 (0.87–2.62)			
			AA	0.030	3	0	4.30 (2.43–5.95)	22.2 (5.72–86.2)			
6	rs1547633		GG	0.678	69	60	16.9 (13.6–18.3)	Ref	2.3×10^{-5}	7.7×10^{-6}	1.00
			GT	0.283	33	26	21.4 (16.2–27.0)	0.76 (0.48–1.21)			
			TT	0.039	3	3	3.58 (3.02–4.30)	29.7 (6.47–136)			
6	rs1570070	*IGF2R*	GG	0.553	66	57	18.2 (15.8–21.4)	Ref	2.2×10^{-5}	0.00010	1.00
			GA	0.388	33	27	16.4 (11.4–17.7)	1.01 (0.63–1.62)			
			AA	0.059	4	4	4.67 (2.17–7.39)	10.5 (3.85–28.9)			
7	rs2711095		GG	0.655	70	59	17.3 (15.9–19.6)	Ref	2.3×10^{-5}	5.0×10^{-5}	1.00
			AG	0.303	30	25	17.3 (11.7–27.0)	1.33 (0.88–2.00)			
			AA	0.042	5	5	5.39 (1.25–9.63)	10.2 (3.8–27.1)			
16	rs4313828	*CNTNAP4*	AA	0.947	99	83	17.4 (15.8–20.4)	Ref	2.2×10^{-5}	8.2×10^{-5}	1.00
			AG	0.050	6	6	7.51 (3.22–9.92)	7.12 (2.87–17.6)			
			GG	0.003	—	—	—	NA			
6	rs894817	*IGF2R*	AA	0.560	65	56	18.3 (15.8–22.3)	Ref	2.8×10^{-5}	0.00012	1.00
			AG	0.379	36	29	16.2 (10.2–17.7)	1.09 (0.69–1.71)			
			GG	0.061	4	4	4.67 (2.17–7.39)	14.3 (4.57–44.9)			
7	rs959494	*SCIN*	AA	0.659	70	56	17.5 (15.9–21.4)	Ref	3.1×10^{-5}	0.00043	1.00
			AG	0.299	30	28	16.0 (8.44–20.3)	1.53 (0.97–2.42)			
			GG	0.042	4	4	5.08 (2.43–9.07)	12.0 (3.97–36.7)			

[a] *p* values were calculated by univariate Cox proportional hazards model.
[b] *p* values were calculated by multivariate Cox proportional hazards model including gender and PS as covariates.
[c] *p* values were adjusted for multiple testing by using the Holm's method.
MST, median survival time; CI, confidence interval; HR, hazard ratio.

in Supplementary Tables 1 (http://links.lww.com/JTO/A43) and 2 (http://links.lww.com/IGC/A24), respectively.

The following PK parameters were measured in this study: AUC PTX (h*/μg/mL), AUC 6-α-hydroxy-PTX (6-α-OH-PTX) (h/μg/ml), AUC C3′-*p*-hydroxy-PTX (3′-*p*-OH-PTX) (h*/μg/mL), AUC 6α-,C3′-*p*-dihydroxy-PTX (diOH-PTX) (h*/μg/mL), AUC Cremophor EL (μl*/h/mL), CL PTX (L/h/m^2). However, no significant association was detected between the PK parameters and the SNPs by a multiple testing correction (data not shown). For reference, we showed the results of association between top 10 SNPs and PK parameters in Supplementary Table 2. This GWAS neither detected a statistically significant association with any of the grade 3/4 adverse reactions (data not shown), probably due to their low incidence, except for neutropenia (Table 2).

DISCUSSION

Cytotoxic chemotherapy continues to be the mainstay for initial treatment of patients with advanced NSCLC. Individualizing chemotherapy to deliver the most active and least toxic agent to each patient could provide an important improvement in patient care.[11] Previous pharmacogenetic studies have identified biomarkers for survival of patients with advanced NSCLC treated with platinum-based chemotherapy.[18–22] Among these are the *XRCC1*, *XRCC3*, and *XPD* genes, which play an important role in DNA repair.[23–28] Similar to previous studies of platinum-based chemotherapy, Gurubhagavatula et al.[18] observed a trend toward decreased survival for patients with variant *XPD* or *XRCC1* genotype and improved survival for patients with variant *XRCC3* genotype.

These genetic polymorphisms were identified by candidate gene approach, which relies on an a priori selection of small numbers of candidate genes based on the existing information or hypothesis. Although successful in several examples, this candidate gene approach may not be able to capture all the genetic factors, which influence a drug response in a complex interplay with multiple unknown as well

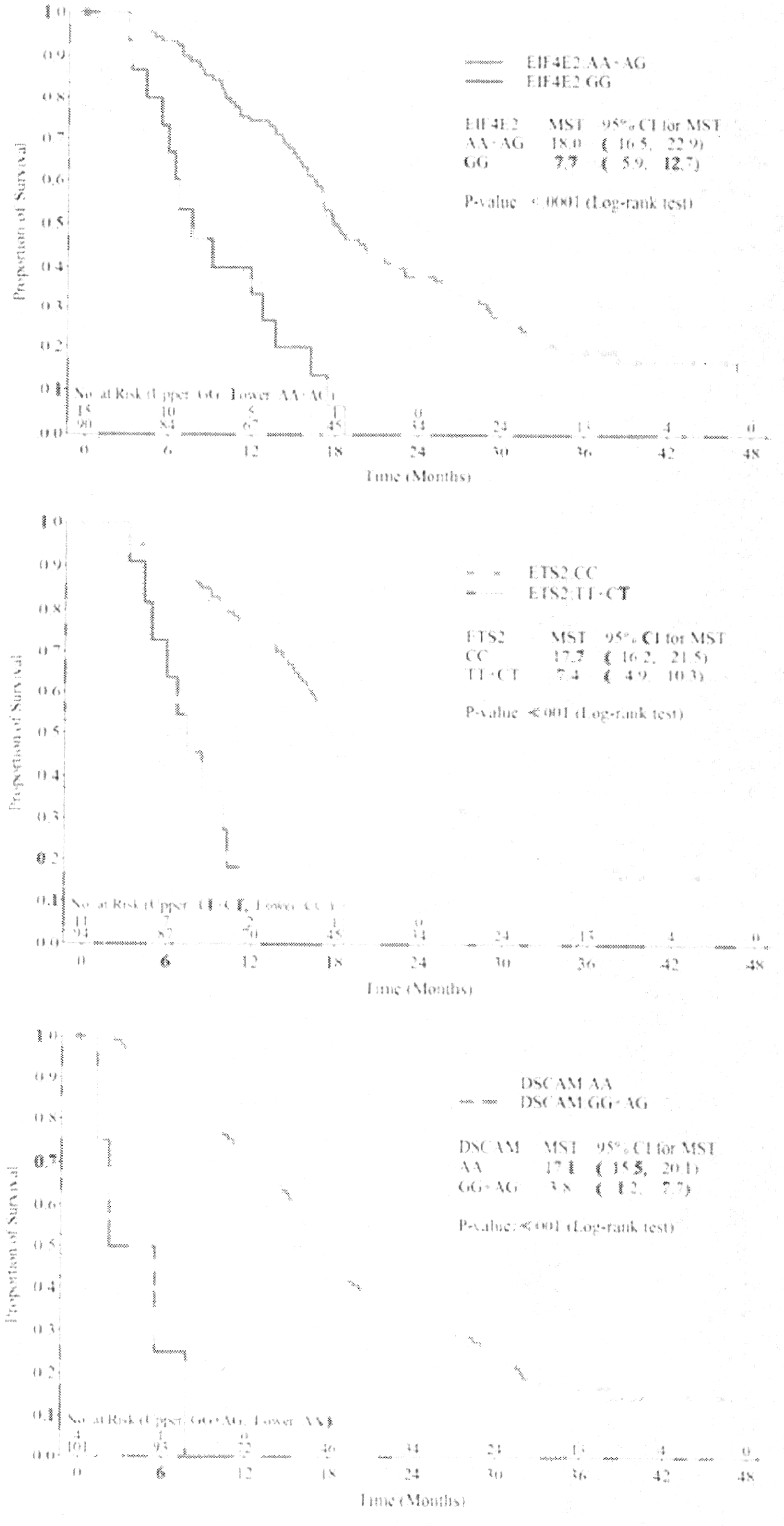

FIGURE 2. Overall survival stratified for the single-nucleotide polymorphism genotype.

as known factors such as disease phenotypes, genetic factors, and the variability in drug target response. GWAS, which makes no assumptions about the genomic location of the causal variants but surveys the whole genome,[29,30] is expected to complement the candidate gene approach. According to our findings from a gene-centric GWAS, three poly-