Table 3 Association between folate and alcohol metabolism genes and breast cancer

Gene name	rs number	Homozygotes of major allele		Heterozygotes				Homozygotes of minor allele			
		Ca	Co	Ca	Co	OR[a] (95% CI)	*P*-value	Ca	Co	OR[a] (95% CI)	*P*-value
MTR	3795708	231	182	252	215	0.94 (0.71–1.25)	0.677	81	81	0.70 (0.48–1.04)	0.075
	1770449	345	296	185	172	0.93 (0.71–1.23)	0.611	36	18	2.21 (1.18–4.16)	0.014
	1050993	344	299	184	172	0.94 (0.71–1.23)	0.640	36	18	2.24 (1.19–4.22)	0.012
MTRR	1801394	295	229	218	210	0.85 (0.65–1.11)	0.228	46	46	0.85 (0.53–1.37)	0.512
	326121	213	184	262	235	0.90 (0.68–1.19)	0.463	91	70	1.01 (0.68–1.50)	0.963
	162049	142	131	285	239	1.19 (0.87–1.63)	0.284	135	118	1.07 (0.75–1.55)	0.704
	10380	409	355	144	123	1.14 (0.84–1.53)	0.398	11	9	1.20 (0.48–3.00)	0.697
	327592	401	346	151	135	1.08 (0.80–1.44)	0.622	11	10	1.05 (0.43–2.58)	0.908
MTHFR	1801133	410	366	144	110	1.23 (0.91–1.68)	0.176	9	11	0.74 (0.29–1.87)	0.520
	1801131	302	258	223	206	0.94 (0.72–1.23)	0.666	38	23	1.57 (0.89–2.79)	0.121
TYMS	16948322	246	234	249	200	1.14 (0.87–1.50)	0.336	65	55	1.27 (0.83–1.94)	0.282
	2298581	242	201	244	205	0.95 (0.72–1.26)	0.723	73	76	0.82 (0.55–1.21)	0.313
ADH1C	4147542	283	250	222	185	1.06 (0.80–1.39)	0.687	45	44	0.90 (0.56–1.43)	0.641
ALDH2	4646778	326	284	191	177	0.91 (0.69–1.21)	0.522	39	24	1.55 (0.88–2.73)	0.131
	671	470	402	79	80	0.85 (0.59–1.22)	0.377	12	4	2.55 (0.76–8.53)	0.129
GSTP1	612020	477	412	82	72	1.06 (0.73–1.52)	0.766	3	6	0.32 (0.08–1.38)	0.126
	1695	307	262	219	196	0.91 (0.70–1.19)	0.492	41	31	1.05 (0.62–1.78)	0.851
NAT1	7845127	277	252	229	199	1.03 (0.79–1.35)	0.826	59	39	1.36 (0.85–2.18)	0.196
NAT2	1041983	167	128	263	245	0.82 (0.60–1.11)	0.191	116	107	0.87 (0.60–1.26)	0.462
	1799930	249	214	231	232	0.82 (0.62–1.08)	0.151	72	37	1.56 (0.99–2.47)	0.058
CYP2E1	2249695	179	162	274	233	1.07 (0.80–1.44)	0.639	96	78	1.05 (0.71–1.56)	0.806
	2031920	400	340	135	115	0.92 (0.68–1.25)	0.595	11	18	0.48 (0.21–1.09)	0.079
DRD2	7117915	254	196	246	223	0.92 (0.70-1.21)	0.543	56	65	0.67 (0.44–1.02)	0.064
	10891556	367	336	174	138	1.21 (0.91–1.61)	0.193	18	13	1.32 (0.62–2.82)	0.476
DRD3	2087017	167	141	276	263	0.84 (0.62–1.14)	0.260	124	87	1.15 (0.79–1.67)	0.465
	167770	315	307	219	150	1.36 (1.03–1.80)	0.029	26	23	0.92 (0.49–1.69)	0.778
SLC6A4	140701	310	287	219	172	1.18 (0.90–1.55)	0.231	38	31	1.33 (0.77–2.29)	0.313

[a] OR adjusted by age (<30, 30–45, 45–60, 60<), BMI (<20, 20–25, 25–30, 30<), smoking, pregnancy and breast feeding, family history of breast cancer, education, and menopausal status

MTHFR is a critical gene in the one-carbon metabolism pathway. Two non-synonymous missense SNPs, C677T (Ala222Val, rs1801133) and A1298C (Glu429Ala, rs1801131), in the coding region were extensively studied. However, the previous association studies on the *MTHFR* polymorphism and breast cancer risk showed inconsistent results. Several studies have reported that the *MTHFR* C677T variants were associated with an increased risk of breast cancer in pre-menopausal women [16, 17] or in those with bilateral breast cancer or combined breast and ovarian cancers [18], but some others showed no association between *MTHFR* C677T and breast cancer [19, 20]. In addition, Sharp et al. [21] reported that *MTHFR* 1298CC genotype and compound heterozygosity (677CT and 1298AC) were associated with a reduced risk of developing breast cancer. In this study, our finding did not support a role of *MTHFR* C677T or *MTHFR* A1298C in modifying breast cancer risk in Thai women.

The 5-methyltetrahydrofolate-homocysteine *S*-methyltransferase (MTR; also called methionine synthase), which is essential for maintaining adequate intracellular folate pools, catalyzes the remethylation of homocysteine to methionine and has influence on DNA methylation as well as on nucleic acid synthesis [22]. Vitamin B12 is a cofactor in this methylation process. MTR is maintained in its active from by methionine synthesis reductase (MTRR). In this study, the homozygote carriers of *MTR* (rs1770449 and rs105099) SNPs were associated with breast cancer with OR = 2.21 (95% CI 1.18–4.16) and OR = 2.24 (95% CI 1.19–4.22), respectively. In addition, the A allele of MTRR (rs162049) seems to be associated with breast cancer among postmenopausal women (OR 1.61, 95% CI 1.07–

Table 4 Association between folate and alcohol metabolism genes and breast cancer stratified by menopausal status

Gene name	Premenopausal women							
	Homozygotes of major allele		Heterozygotes			Homozygotes of minor allele		
	Ca	Co	Ca	Co	OR[a] (95% CI)	Ca	Co	OR[a] (95% CI)
MTRR (rs162049)	69	39	119	89	0.79 (0.48–1.32) $P = 0.372$	51	46	0.74 (0.41–1.35) $P = 0.328$
NAT2 (rs1799930)	109	82	89	83	0.82 (0.53–1.26) $P = 0.370$	35	9	2.70 (1.20–6.07) $P = 0.017$
DRD2 (rs10891556)	147	122	83	46	1.62 (1.03–2.56) $P = 0.038$	6	6	0.94 (0.28–3.16) $P = 0.925$
DRD3 (rs167770)	142	106	80	54	1.13 (0.72–1.76) $P = 0.593$	14	13	0.75 (0.32–1.74) $P = 0.499$
SLC6A4 (rs140701)	128	109	97	56	1.56 (1.01–2.41) $P = 0.046$	15	11	1.13 (0.47–2.74) $P = 0.785$
Gene name	Postmenopausal women							
	Homozygotes of major allele		Heterozygotes			Homozygotes of minor allele		
	Ca	Co	Ca	Co	OR[a] (95% CI)	Ca	Co	OR[a] (95% CI)
MTRR (rs162049)	71	91	166	150	1.61 (1.07–2.44) $P = 0.024$	82	72	1.43 (0.14–2.31) $P = 0.889$
NAT2 (rs1799930)	138	132	141	148	0.81 (0.56–1.16) $P = 0.240$	36	28	1.10 (0.76–1.96) $P = 0.612$
DRD2 (rs10891556)	216	213	91	92	0.98 (0.67–1.43) $P = 0.912$	12	7	1.50 (0.44–4.14) $P = 0.543$
DRD3 (rs167770)	171	200	137	96	1.59 (1.11–2.28) $P = 0.011$	12	10	1.16 (0.75–2.89) $P = 0.468$
SLC6A4 (rs140701)	180	178	120	115	0.93 (0.65–1.33) $P = 0.687$	23	20	1.46 (0.30–2.96) $P = 0.720$

[a] OR adjusted by age (<30, 30–45, 45–60, 60<), BMI (<20, 20–25, 25–30, 30<), smoking, pregnancy and breast feeding, family history of breast cancer, and education

2.44). However, these SNPs have never been reported with breast cancer before and their functional effect on enzymatic activities remain unknown.

Alcohol drinking is among the non-hormonal risk factors for breast cancer (although there may be an indirect relationship). Genetic susceptibility for ethanol metabolism can affect breast cancer, and several genes are known to be involved in this complex pathway. Our study did not find any affect of *ADH1C*, *ALDH2*, *CYP2E1*, *GSTP1*, and *NAT1* polymorphisms on breast cancer risk. However, the homozygotes of minor allele of *NAT2* (rs1799930) SNP was associated with an increased risk of breast cancer among premenopausal (OR 2.70, 95% CI 1.20–6.07), although the number of the subjects were limited. Lu et al. [23] showed an interaction effect in bladder cancer between *NAT2* polymorphism and alcohol drinking. In addition, Rodrigo et al. [24] reported that NAT2 activity may be a factor that determines the risk of developing alcoholic liver disease. In this study, the genotype distribution of the *NAT2* (rs1799930) polymorphism departed from Hardy–Weinberg Equilibrium, and the interpretation of the result should be warranted.

Furthermore, many studies suggested that the intensity of drinking (drink per day) has more effect on risk for breast cancer than recent alcohol or duration of drinking [25, 26]. We, therefore, investigated the SNPs of *DRD2*, *DRD3*, and *SLC6A4*, which are implicated in drinking behavior. In animal studies, alcohol can stimulate dopaminergic neurons in the ventral tegmental area [27, 28], and the density of dopamine D2 receptors in the limbic system is lower in alcohol-preferring rats than in non-preferring rats [29, 30]. Likewise, the number of striatal dopamine D2 receptors is less in alcohol-preferring humans than in healthy control subjects. The A1 polymorphism of *DRD2 Taq*I A loci has been considered as a risk factor for alcohol dependence [31, 32], but the association between alcoholism and the *DRD2* gene remains equivocal in many studies [33–35]. In our study, an

Table 5 Association of selected SNPs and breast cancer risk by alcohol, BMI, and physical activity

Variable	*MTR* (rs1770449) (case/control)		OR[a] (95% CI)		*MTR* (rs1050993) (case/control)	
	AA	AG/GG	AA	AG/GG	GG	GA/GG
Alcohol						
Yes	27/14	18/12	1	1.198 (0.28–5.13)	27/15	17/12
No	203/178	318/282	1	1.053 (0.80–1.38)	317/284	203/178
P for interaction				$P = 0.64$		
BMI (kg/m^2)						
<25	217/224	142/146	1	1.033 (0.75–1.41)	216/228	141/145
≥25	128/72	79/44	1	1.042 (0.63–1.72)	128/71	79/45
P for interaction				$P = 0.97$		
OC use						
Yes	157/121	99/76	1	1.054 (0.71–1.56)	157/123	99/76
No	187/171	122/114	1	1.007 (0.71–1.44)	186/172	121/114
P for interaction				$P = 0.75$		

Variable	OR (95% CI)		*DRD3* (rs167770) (case/contol)		OR (95% CI)	
	GG	GA/GG	AA	AG/GG	AA	AG/GG
Alcohol						
Yes	1	1.339 (0.32–5.53)	27/16	17/11	1	0.518 (0.12–2.23)
No	1	1.059 (0.81–1.39)	288/291	228/162		1.355 (1.03–1.78)
P for interaction		$P = 0.69$				$P = 0.24$
BMI (kg/m^2)						
<25	1	1.063 (0.78–1.45)	210/235	143/129	1	1.204 (0.88–1.65)
≥25	1	1.002 (0.61–1.65)	105/72	102/44	1	0.253 (0.06–1.02)
P for interaction		$P = 0.86$				$P = 0.39$
OC use						
Yes	1	1.026 (0.72–1.47)	141/127	115/67	1	1.111 (0.77–1.59)
No	1	1.054 (0.71–1.57)	174/178	129/104	1	1.56 (1.05–2.33)
P for interaction		$P = 0.81$				$P = 0.21$

[a] OR adjusted by age(<30, 30–45, 45–60, 60<), BMI (<20, 20–25, 25–30, 30<), smoking, menopausal status, family history of breast cancer, and education

Table 6 Association of *DRD3* and *MTRR* polymorphisms with the risk of breast cancer

Genotype	*DRD3* (AA)				*DRD3*(AG/GG)				*P* for interaction
	Cases	Controls	OR[a]	95% CI	Cases	Controls	OR	95% CI	
MTR (rs1770449)									
AA	184	188	1		155	102	1		
AG	108	103	1.07	0.74–1.53	74	67	0.72	0.46–1.13	0.096
GG	20	14	1.93	0.89–4.18	16	4	3.78	1.04–13.6	0.25
MTR (rs1050993)									
GG	183	191	1		155	102	1		
GA	108	102	1.09	0.76–1.56	74	66	0.73	0.46–1.15	0.10
AA	20	14	1.96	0.91–4.24	16	4	3.77	1.04–13.6	0.26

[a] OR adjusted by age (<30, 30–45, 45–60, 60<), BMI (<20, 20–25, 25–30, 30<), smoking, menopausal status, family history of breast cancer, and education

association was suggested between the *DRD3* SNP (rs167770) in an overall or in postmenopausal breast cancer (OR 1.36, 95% CI 1.03–1.80 and OR 1.59, 95% CI 1.11–2.28, respectively). The *DRD2* (rs10891556) polymorphism was associated with an increased risk among premenopausal women (OR 1.62, 95% CI 1.03–2.56). The propensity for severe drinking has been also hypothesized to be regulated by differential expression of serotonin transporter gene *SLC6A4* [36]. Our result suggested that *SCL6A4* (rs140701) was associated with an increased risk of breast cancer among premenopausal (OR 1.56, 95% CI 1.01–2.41). As the alcohol consumption was relatively low among Thai women in our study, it may have reduced the ability to detect a modifying effect of ethanol on the association of these genes with breast cancer. The current knowledge on genetic polymorphisms related to drinking behavior is far from sufficient. The possible associations between the genetic polymorphisms implicated in drinking behavior and breast cancer risk need to be confirmed in a population with a higher prevalence of alcohol drinking than Thai women.

Certain limitations of this study should be noted. First, data were not available on detailed dietary intake of folate, plasma or erythrocyte folate levels and its precursors or metabolites such as homocysteine, limiting further examination of the gene–nutrient interactions in breast carcinogenesis. In this study, a food frequency questionnaire was administered by trained interviewers to assess dietary and alcohol intake. Unfortunately, we could not calculate the total folate in individual intake due to lack of the standard food composition in Thailand. Second, like most other case–control studies, this study may suffer from recall bias. Third, the statistical power of our study was limited in the stratified analyses because of the small sample size of the subgroups. For instance, if the simple but conservative Bonferroni correction for multiple testing is applied to our data set, none of the SNPs remained statistically significant. Although this study reports a systematic survey of genetic polymorphisms on the folate and alcohol metabolic pathways on breast cancer in Thailand for the first time, it is primarily for a hypothesis generation, and the findings need to be validated in further studies with larger sample size or in meta-analyses, which is also aimed in our laboratory for years to come.

In conclusion, our study has provided some new evidence that either folate metabolism genes or alcohol metabolism and behavior gene polymorphisms may contributes to the etiology of breast cancer among Thai women. Studies should be extended to cover SNPs of other important one-carbon or alcohol metabolism genes, and ascertainment of high quality folate intake information is expected to further elucidate gene–gene and gene–environment interactions in susceptibility of breast cancer.

Acknowledgments We express our sincere thanks to all patients, recruited control subjects, doctors, nurses, and paramedical personnel of National Cancer Institute, Thailand for their kind participation. This work was supported by the program for promotion of Fundamental Studies in Health Sciences of the National Institute of Biomedical Innovation (NiBio 05-41), Japan. S. Sangrajrang is an awardee of a fellowship from Foundation for Promotion of Cancer Research.

References

1. Sriplung H, Sontipong S, Martin N, Wiangnon S, Vootiprux V, Cheirsilpa A, Kanchanabat C, Khuhaprema T (2005) Cancer incidence in Thailand, 1995–1997. Asian Pac J Cancer Prev 6(3):276–281
2. Stover PJ (2004) Physiology of folate and vitamin B12 in health and disease. Nutr Rev 62(6Pt2):S3–S12
3. Kim YI (1999) Folate and carcinogenesis: evidence, mechanisms and implications. J Nutr Biochem 10(2):66–88
4. Longnecker MP (1994) Alcoholic beverage consumption in relation to risk of breast cancer: meta-analysis and review. Cancer Causes Control 5(1):73–82
5. Smith-Warner SA, Spiegelman D, Yaun SS, van den Brandt PA, Folsom AR, Goldbohm RA, Graham S, Holmberg L, Howe GR, Marshall JR, Miller AB, Potter JD, Speizer FE, Willett WC, Wolk A, Hunter DJ (1998) Alcohol and breast cancer in women: a pooled analysis of cohort studies. JAMA 279(7):535–540
6. Singletary KW, Gapstur SM (2001) Alcohol and breast cancer: review of epidemiologic and experimental evidence and potential mechanisms. JAMA 286(17):2143–2151
7. Purohit V (1998) Moderate alcohol consumption and estrogen levels in postmenopausal women: a review. Alcohol Clin Exp Res 22(5):994–997
8. Thomas DB (1995) Alcohol as a cause of cancer. Environ Health Perspect 103(Suppl. 8):153–160
9. Feron VJ, Til HP, de Vrijer F, Woutersen RA, Cassee FR, van Bladeren PJ (1991) Aldehydes: occurrence, carcinogenic potential, mechanism of action and risk assessment. Mutat Res 259(3–4):363–385
10. Halsted CH, Villanueva JA, Devlin AM, Chandler CJ (2002) Metabolic interactions of alcohol and folate. J Nutr 132(Suppl. 8):2367S–2372S
11. Zhang S, Willett WC, Selhub J, Hunter DJ, Giovannucci EL, Holmes MD, Colditz GA, Hankinson SE (2003) Plasma folate, vitamin B6, vitamin B12, homocysteine, and risk of breast cancer. J Natl Cancer Inst 95(5):373–380
12. Rohan TE, Jain MG, Howe GR, Miller AB (2000) Dietary folate consumption and breast cancer risk. J Natl Cancer Inst 92(3):266–269
13. Sellers TA, Kushi LH, Cerhan JR, Vierkant RA, Gapstur SM, Vachon CM, Olson JE, Therneau TM, Folsom AR (2001) Dietary folate intake, alcohol, and risk of breast cancer in a prospective study of postmenopausal women. Epidemiology 12(4):420–428
14. Negri E, La Vecchia C, Franceschi S (2000) Re: dietary folate consumption and breast cancer risk. J Natl Cancer Inst 92(15):1270–1271
15. Yoshimura K, Hanaoka T, Ohnami S, Ohnami S, Kohno T, Liu Y, Yoshida T, Sakamoto H, Tsugane S (2003) Allele frequencies of single nucleotide polymorphisms (SNPs) in 40 candidate genes for gene-environment studies on cancer: data from population-based Japanese random samples. J Hum Genet 48(12):654–658
16. Ergul E, Sazci A, Utkan Z, Canturk NZ (2003) Polymorphisms in the MTHFR gene are associated with breast cancer. Tumor Biol 24(6):286–290

17. Semenza JC, Delfino RJ, Ziogas A, Anton-Culver H (2003) Breast cancer risk and methylenetetrahydrofolate reductase polymorphism. Breast Cancer Res Treat 77(3):217–223
18. Gershoni-Baruch R, Dagan E, Israeli D, Kasinetz L, Kadouri E, Friedman E (2000) Association of the C677T polymorphism in the MTHFR gene with breast and/or ovarian cancer risk in Jewish women. Eur J Cancer 36(18):2313–2316
19. Yu CP, Wu MH, Chou YC, Yang T, You SL, Chen CJ, Sun CA (2007) Breast cancer risk associated with multigenotypic polymorphisms in folate-metabolizing genes: a nested case–control study in Taiwan. Anticancer Res 27(3B):1727–1732
20. Le Marchand L, Haiman CA, Wilkens LR, Kolonel LN, Henderson BE (2004) MTHFR polymorphisms, diet, HRT, and breast cancer risk: the multiethnic cohort study. Cancer Epidemiol Biomarkers Prev 13(12):2071–2077
21. Sharp L, Little J, Schofield AC, Pavlidou E, Cotton SC, Miedzybrodzka Z, Baird JO, Haites NE, Heys SD, Grubb DA (2002) Folate and breast cancer: the role of polymorphisms in methylenetetrahydrofolate reductase (MTHFR). Cancer Lett 181(1):65–71
22. Linnebank M, Schmidt S, Kölsch H, Linnebank A, Heun R, Schmidt-Wolf IG, Glasmacher A, Fliessbach K, Klockgether T, Schlegel U, Pels H (2004) The methionine synthase polymorphism D919G alters susceptibility to primary central nervous system lymphoma. Br J Cancer 90(10):1969–1971
23. Lu CM, Chung MC, Huang CH, Ko YC (2005) Interaction effect in bladder cancer between N-acetyltransferase 2 genotype and alcohol drinking. Urol Int 75(4):360–364
24. Rodrigo L, Alvarez V, Rodriguez M, Pérez R, Alvarez R, Coto E (1999) N-acetyltransferase-2, glutathione S-transferase M1, alcohol dehydrogenase, and cytochrome P450IIE1 genotypes in alcoholic liver cirrhosis: a case–control study. Scand J Gastroenterol 34(3):303–307
25. Hamajima N, Hirose K, Tajima K et al (2002) Alcohol, tobacco and breast cancer—collaborative reanalysis of individual data from 53 epidemiological studies, including 58, 515 women with breast cancer and 95, 067 women without the disease. Br J Cancer 87(11):1234–1245
26. Bowlin SJ, Leske MC, Varma A, Nasca P, Weinstein A, Caplan L (1997) Breast cancer risk and alcohol consumption: results from a large case–control study. Int J Epidemiol 26(5):915–923
27. Brodie MS, Shefner SA, Dunmiddie TV (1990) Ethanol increases the firing rate of dopamine neurons of the rat ventral tegmental area in vitro. Brain Res 508(1):65–69
28. Brodie MS (2002) Increased ethanol excitation of dopaminergic neurons of the ventral tegmental area after chronic ethanol treatment. Alcohol Clin Exp Res 26(7):1024–1030
29. McBride WJ, Chernet E, Dyr W, Lumeng L, Li TK (1993) Densities of dopamine D2 receptors are reduced in CNS regions of alcohol-preferring P rats. Alcohol 10(5):387–390
30. Stefanini E, Frau M, Garau MG, Garau B, Fadda F, Gessa GL (1992) Alcohol-preferring rats have fewer dopamine D2 receptors in the limbic system. Alcohol Alcohol 27(2):127–130
31. Noble EP, Blum K, Ritchie T, Montgomery A, Sheridan PJ (1991) Allelic association of the D2 dopamine receptor gene with receptor-binding characteristics in alcoholism. Arch Gen Psychiatry 48(7):648–654
32. Reich T, Hinrichs A, Culverhouse R, Bierut L (1999) Genetic studies of alcoholism and substance dependence. Am J Hum Genet 65(3):599–605
33. Blum K, Noble EP, Sheridan PJ, Montgomery A, Ritchie T, Jagadeeswaran P, Nogami H, Briggs AH, Cohn JB (1990) Allelic association of human dopamine D2 receptor gene in alcoholism. JAMA 263(15):2055–2060
34. Bolos AM, Dean M, Lucas-Derse S, Ramsburg M, Brown GL, Goldman D (1990) Population and pedigree studies reveal a lack of association between the dopamine D2 receptor gene and alcoholism. JAMA 264(24):3156–3160
35. Huang SY, Lin WW, Ko HC, Lee JF, Wang TJ, Chou YH, Yin SJ, Lu RB (2004) Possible interaction of alcohol dehydrogenase and aldehyde dehydrogenase genes with the dopamine D2 receptor gene in anxiety-depressive alcohol dependence. Alcohol Clin Exp Res 28(3):374–384
36. Philibert RA, Sandhu H, Hollenbeck N, Gunter T, Adams W, Madan A (2008) The relationship of 5HTT (SLC6A4) methylation and genotype on mRNA expression and liability to major depression and alcohol dependence in subjects from the Iowa Adoption Studies. Am J Med Genet B 147B(5):543–549

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.

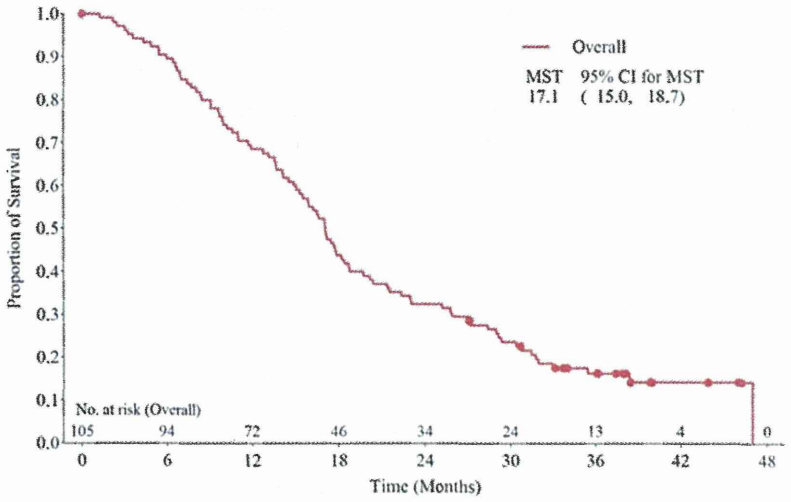

FIGURE 1. Kaplan-Meier plot for overall survival.

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.

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.

(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.

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

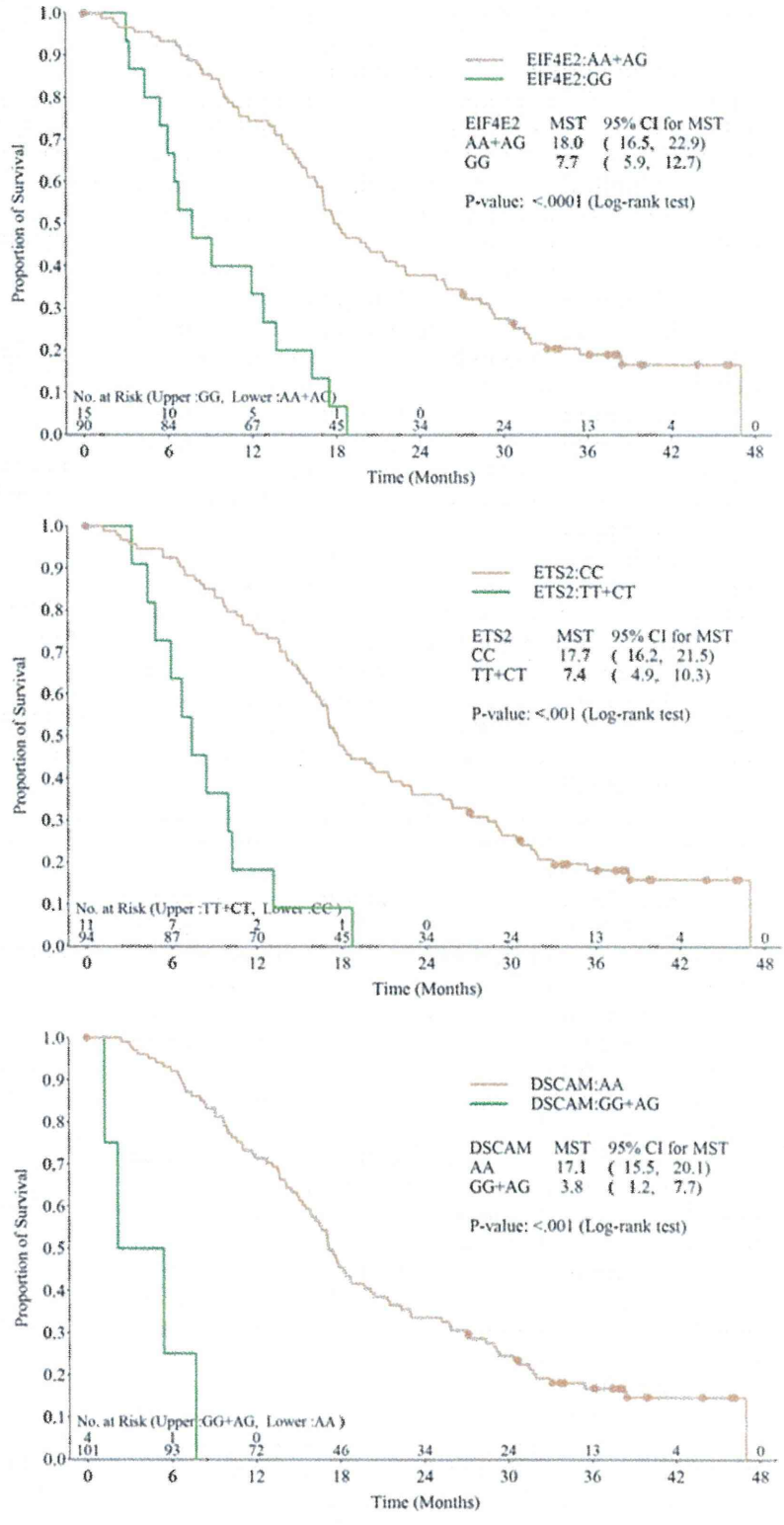

FIGURE 2. Overall survival stratified for the single-nucleotide polymorphism genotype.

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-

morphisms were associated with shortened OS in advanced NSCLC with CBDCA and PTX. The three SNPs have not been previously investigated for an association with NSCLC risk or drug response. On the other hand, the SNPs implicated in the prognosis of NSCLC by the previous candidate gene approach[18] were not detected in the GWAS, because the Human-1 BeadChip does not harbor the identical SNPs analyzed before and/or their p values were not sufficiently small in the context of the genome scan.

The first candidate SNP for the OS association, rs1656402, is in the third intron of the gene, *EIF4E2*, encoding for the translational factor eukaryotic initiation factor 4E, which is a central component in the initiation and regulation of translation in eukaryotic cells. Through its interaction with the 5′ cap structure of mRNA, eIF4E functions to recruit mRNAs to the ribosome.[31–34] Prototypical eIF4E-2 is expressed ubiquitously,[33,35] but in metastatic tumors, its expression was increased,[36] suggesting that eIF4E-2 plays an active role in the prognosis of NSCLC.

The second candidate SNP is located at the 4321 bp upstream of the *ETS2* gene. The Ets family of transcription factors includes important downstream targets in cellular transformation. For instance, alteration of Ets activity has been found to reverse the transformed phenotype of ras-transfected mouse fibroblasts and of several human tumor cell lines. It has been reported that Ets factor activity can strongly influence the transformed and invasive phenotype of a human prostate tumor cell line.[37]

The third candidate rs9981861 is in the 31st intron of the 33-exon *DSCAM* gene, which encodes Down syndrome cell adhesion molecule, a member of the immunoglobulin superfamily. The gene was cloned from the Down syndrome region on chromosome 21q22 and found to be expressed widely in the developing nervous system.[38] Mouse *DSCAM* has been shown to mediate arborization of neurite processes and spacing of neuronal cell bodies.[39,40] Expression of the *DSCAM* gene has been upregulated in small cell lung cancer compared with NSCLC.[41]

Because a GWAS is based on a linkage disequilibrium (LD) mapping of a disease locus by use of SNPs as markers, the particular SNPs per se identified in this study may not be functionally responsible for the observed effect on survival time. In fact, LD maps drawn by the HapMap data around the three SNPs indicate that at least the SNPs of the *EIF4E2* and *ETS2* genes are embedded in extended LD blocks (Supplementary Figure 1, http://links.lww.com/IGC/A25); it may be then difficult to narrow down the regions of interest further for these SNPs by statistical genetics alone, at least in the Asian population.

In summary, a hypothesis-free GWAS detected previously unrecognized associations between polymorphisms of the three genes and shortened OS in advanced NSCLC treated with CBDCA and PTX. Additionally, these three SNPs on the three genes were significant after a multiple testing adjustment. In considering a multiple testing problem, we assume the existence of about 10,000 linkage disequilibrium blocks within 100,000 gene-centric SNPs, which are concentrated in about 2% of the human genome (i.e., average interval of two SNPs is 600 bp). It follows that the p value cutoff is set at 5.0×10^{-6} if the Bonferroni correction is applied. However, in the first screening, such correction for a multiple testing is often too conservative, failing to detect many drug-response SNPs; therefore, we showed top 10 SNPs in Table 4. In addition, to facilitate the second screening or replication studies by other investigators, statistics of association between OS, PK parameters, toxicity, and all SNPs analyzed in this study are available at Genome Medicine Database of Japan (http://gemdbj.nibio.go.jp).

The ultimate goal of this work is better clinical management of patients after the assessment of genotype risk on OS. To this end, however, we need to identify genetic polymorphisms that can differentiate patients' response and outcome to different chemotherapeutic agents. Although our work may contribute as the first step to establish such a predictive factor, especially the survival-related SNPs that also influence pharmacokinetics, the current single-arm prospective study does not provide definite evidence of pharmacogenomic profiling for a platinum-based chemotherapy. Several targeted therapies for NSCLC are in clinical development, and it is hoped that this line of pharmacogenetic studies will eventually help clinicians to choose platinum or nonplatinum doublets as the first-line regimen, for instance. Further studies of NSCLC would stratify patients according to the SNP status to tailor treatment to individual patients. The results of a single association study should be validated by independent studies by other investigators as well as biologic functional analyses.

ACKNOWLEDGMENTS

Supported by the Program for the Promotion of Fundamental Studies in Health Sciences from National Institute of Biomedical Innovation (ID 05-41).

REFERENCES

1. Cancer Statistics in Japan 2008: The Editorial Board of the Cancer Statistics in Japan. Tokyo, Japan: Foundation for Promotion of Cancer Research 2008. Available at: http://www.fpcr.or.jp/publication/statistics.html. Accessed March 3, 2010.
2. Non-Small Cell Lung Cancer Collaborative Group. Chemotherapy in non-small cell lung cancer: a meta-analysis using updated data on individual patients from 52 randomized clinical trials. *BMJ* 1995;311:899–909.
3. Fukuoka M, Niitani H, Suzuki A, et al. A phase II study of CPT-11, a new derivative of camptothecin, for previously untreated non-small-cell lung cancer. *J Clin Oncol* 1992;10:16–20.
4. Rowinsky EK, Donehower RC. Paclitaxel (taxol). *N Engl J Med* 1995;332:1004–1014.
5. Gelmon K. The taxoids: paclitaxel and docetaxel. *Lancet* 1994;344:1267–1272.
6. Hertel LW, Border GB, Kroin JS, et al. Evaluation of the antitumor activity of gemcitabine. *Cancer Res* 1990;50:4417–4422.
7. Binet S, Fellous A, Lataste H, et al. Biochemical effects of navelbine on tubulin and associated proteins. *Semin Oncol* 1989;16:9–14.
8. Petris L, Crino L, Scagliotti GV, et al. Treatment of advanced non-small cell lung cancer. *Ann Oncol* 2006;17(Suppl 2):ii36–ii41.
9. Kubota K, Kawahara M, Ogawara M, et al. Vinorelbine plus gemcitabine followed by docetaxel versus carboplatin plus paclitaxel in patients with advanced non-small-cell lung cancer: a randomised, open-label, phase III study. *Lancet Oncol* 2008;9:1135–1142.
10. Bepler G. Using translational research to tailor the use of chemotherapy in the treatment of NSCLC. *Lung Cancer* 2005;50(Suppl 1):S13–S14.

11. Rosell R, Cobo M, Isla D, et al. Applications of genomics in NSCLC. *Lung Cancer* 2005;50:S33–S40.
12. Nakajima Y, Yoshitani T, Fukushima-Uesaka H, et al. Impact of the haplotype CYP3A4*16B harboring the Thr185Ser substitution on paclitaxel metabolism in Japanese patients with cancer. *Clin Pharmacol Ther* 2006;80:179–191.
13. Saito Y, Katori N, Soyama A, et al. CYP2C8 haplotype structures and their influence on pharmacokinetics of paclitaxel in a Japanese population. *Pharmacogenet Genomics* 2007;17:461–471.
14. Therasse P, Arbuck SG, Eisenhauer EA, et al. New guidelines to evaluate the response to treatment in solid tumors. European Organization for Research and Treatment of Cancer, National Cancer Institute of the United States, National Cancer Institute of Canada. *J Natl Cancer Inst* 2000;92:205–216.
15. Cox, DR. Regression models and life tables. *J R Stat Soc* 1972;34:187–220.
16. Kalbfleisch JD, Prentice RL. The Statistical Analysis of Failure Time Data. New York, NY: John Wiley and Sons, 1980.
17. Holm S. A simple sequentially rejective multiple test procedure. *Scand J Stat* 1979;6:65–70.
18. Gurubhagavatula S, Liu G, Park S, et al. XPD and XRCC1 genetic polymorphisms are prognostic factors in advanced non-small-cell lung cancer patients treated with platinum chemotherapy. *J Clin Oncol* 2004;22:2594–2601.
19. Isla D, Sarries C, Rosell R, et al. Single nucleotide polymorphisms and outcome in docetaxel-cisplatin-treated advanced non-small-cell lung cancer. *Ann Oncol* 2004;15:1194–1203.
20. Ryu JS, Hong YC, Han HS, et al. Association between polymorphisms of ERCC1 and XPD and survival in non-small cell lung cancer patients treated with cisplatin combination chemotherapy. *Lung Cancer* 2004; 44:311–316.
21. de las Penas R, Sanchez-Ronco M, Alberola V, et al. Polymorphisms in DNA repair genes modulate survival in cisplatin/gemcitabine-treated non-small-cell lung cancer patients. *Ann Oncol* 2006;17:668–675.
22. Booton R, Ward T, Heighway J, et al. Xeroderma pigmentosum group D haplotype predicts for response, survival, and toxicity after platinum-based chemotherapy in advanced nonsmall cell lung cancer. *Cancer* 2006;106:2421–2427.
23. Spitz MR, Wu X, Wang Y, et al. Modulation of nucleotide excision repair capacity by XPD polymorphisms in lung cancer patients. *Cancer Res* 2001;61:1354–1357.
24. Duell EJ, Wiencke JK, Cheng TJ, et al. Polymorphisms in the DNA repair genes XRCC1 and ERCC2 and biomarkers of DNA damage in human blood mononuclear cells. *Carcinogenesis* 2000;21:965–971.
25. Matullo G, Palli D, Peluso M, et al. XRCC1, XRCC3, XPD gene polymorphisms, smoking and (32)P-DNA adducts in a sample of healthy subjects. *Carcinogenesis* 2001;22:1437–1445.
26. Bosken CH, Wei Q, Amos CI, et al. An analysis of DNA repair as a determinant of survival in patients with non-small-cell lung cancer. *J Natl Cancer Inst* 2002;94:1091–1099.
27. Wei Q, Wang X, Shen H. DNA repair gene XPD polymorphism and lung cancer risk: a meta-analysis. *Lung Cancer* 2004;46:1–10.
28. Chen J, Larochelle S, Li X, et al. Xpd/Ercc2 regulates CAK activity and mitotic progression. *Nature* 2003;424:228–232.
29. Hirschhorn JN, Daly MJ. Genome-wide association studies for common diseases and complex traits. *Nat Rev Genet* 2005;6:95–108.
30. Nordborg M, Tavaré B. Linkage disequilibrium: what history has to tell us. *Trends Genet* 2002;18:83–90.
31. Gingras AC, Raught B, Sonenberg N. eIF4 initiation factors: effectors of mRNA recruitment to ribosomes and regulators of translation. *Annu Rev Biochem* 1999;68:913–963.
32. Gross JD, Moerke NJ, von der Haar T, et al. Ribosome loading onto the mRNA cap is driven by conformational coupling between eIF4G and eIF4E. *Cell* 2003;115:739–750.
33. Joshi B, Cameron A, Jagus R. Characterization of mammalian eIF4E-family members. *Eur J Biochem* 2004;271:2189–2203.
34. Okumura F, Zou W, Zhang DE. ISG15 modification of the eIF4E cognate 4EHP enhances cap structure-binding activity of 4EHP. *Genes Dev* 2007;21:255–260.
35. Rom E, Kim HC, Gingras AC, et al. Cloning and characterization of 4EHP, a novel mammalian eIF4E-related cap-binding protein. *J Biol Chem* 1998;273:13104–13109.
36. Ramaswamy S, Ross KN, Lander ES, et al. A molecular signature of metastasis in primary solid tumors. *Nat Genet* 2003;33:49–54.
37. Foos G, Hauser CA. Altered Ets transcription factor activity in prostate tumor cells inhibits anchorage-independent growth, survival, and invasiveness. *Oncogene* 2000;19:5507–5516.
38. Yamakawa K, Huo Y-K, Haendel MA, et al. DSCAM: a novel member of the immunoglobulin superfamily maps in a Down syndrome region and is involved in the development of the nervous system. *Hum Mol Genet* 1998;7:227–237.
39. Wojtowicz WM, Flanagan JJ, Millard S, et al. Alternative splicing of *Drosophila* Dscam generates axon guidance receptors that exhibit isoform-specific homophilic binding. *Cell* 2004;118:619–633.
40. Fuerst PG, Koizumi A, Masland RH, et al. Neurite arborization and mosaic spacing in the mouse retina require DSCAM. *Nature* 2008;451: 470–474.
41. Coe BP, Lockwood WW, Girard L, et al. Differential disruption of cell cycle pathways in small cell and non-small cell lung cancer. *Br J Cancer* 2006;94:1927–1935.

A Functional Single Nucleotide Polymorphism in *Mucin 1*, at Chromosome 1q22, Determines Susceptibility to Diffuse-Type Gastric Cancer

NORIHISA SAEKI,[*] AKIRA SAITO,[*,‡] IL JU CHOI,[§] KEITARO MATSUO,[‖] SUMIKO OHNAMI,[*] HIROHIKO TOTSUKA,[*,¶] SUENORI CHIKU,[*,#] AYA KUCHIBA,[*] YEON-SU LEE,[§] KYONG-AH YOON,[§] MYEONG-CHERL KOOK,[§] SOOK RYUN PARK,[§] YOUNG-WOO KIM,[§] HIDEO TANAKA,[‖] KAZUO TAJIMA,[‖] HIROSHI HIROSE,[**] FUMIHIKO TANIOKA,[‡‡] YOSHIHIRO MATSUNO,[§§] HARUHIKO SUGIMURA,[‖‖] SHUNJI KATO,[¶¶] TSUNEYA NAKAMURA,[##] TOMOHIRO NISHINA,[***] WATARU YASUI,[‡‡‡] KAZUHIKO AOYAGI,[*] HIROKI SASAKI,[*] KAZUYOSHI YANAGIHARA,[§§§] HITOSHI KATAI,[‖‖‖] TADAKAZU SHIMODA,[¶¶¶] TERUHIKO YOSHIDA,[*] YUSUKE NAKAMURA,[###,****] SETSUO HIROHASHI,[****] and HIROMI SAKAMOTO[*]

[*]Genetics Division, National Cancer Center Research Institute, Tokyo; [‡]Statistical Genetics Analysis Division, StaGen Co, Ltd, Tokyo, Japan; [§]Research Institute and Hospital, National Cancer Center, Gyeonggi-do, Korea; [‖]Division of Epidemiology and Prevention, Aichi Cancer Center Research Institute, Nagoya; [¶]Bioinfomatics Group, Research and Development Center, Solution Division 4, Hitachi Government and Public Corporation System Engineering Ltd, Tokyo; [#]Science Solutions Division, Mizuho Information and Research Institute, Inc, Tokyo; [**]Department of Internal Medicine, Keio University School of Medicine, Tokyo; [‡‡]Iwata City Hospital, Shizuoka; [§§]Department of Surgical Pathology, Hokkaido University Hospital, Sapporo; [‖‖]First Department of Pathology, Hamamatsu University School of Medicine, Shizuoka; [¶¶]Department of Surgery, Nippon Medical School Hospital, Tokyo; [##]Department of Endoscopy, Aichi Cancer Center Hospital, Nagoya; [***]Department of Internal Medicine, Shikoku Cancer Center, Ehime; [‡‡‡]Department of Molecular Pathology, Hiroshima University Graduate School of Biomedical Sciences, Hiroshima; [§§§]Department of Life Sciences, Yasuda Women's University Faculty of Pharmacy, Hiroshima; [‖‖‖]Department of Surgical Oncology, National Cancer Center Hospital, Tokyo; [¶¶¶]Center for Cancer Control and Information Services, National Cancer Center, Tokyo; [###]Human Genome Center, Institute of Medical Science, University of Tokyo, Tokyo; and [****]National Cancer Center Research Institute, Tokyo, Japan

BACKGROUND & AIMS: Two major types of gastric cancer, intestinal and diffuse, develop through distinct mechanisms; the diffuse type is considered to be more influenced by genetic factors, although the mechanism is unknown. Our previous genome-wide association study associated 3 single nucleotide polymorphisms (SNPs) with diffuse-type gastric cancer (DGC); 1 was a functional SNP (rs2294008) in *prostate stem cell antigen* (*PSCA*), but the loci of the other 2 were not investigated. **METHODS:** We performed high-density mapping to explore a linkage disequilibrium status of the 2 SNPs at chromosome 1q22. A DGC case-control study was conducted using DNA from 606 cases and 1264 controls (all Japanese individuals) and validated using DNA from Japanese (304 cases, 1465 controls) and Korean (452 cases, 372 controls) individuals. The effects of SNPs on function were analyzed by reporter assays and analyses of splice variants. **RESULTS:** A region of a strong linkage disequilibrium with the 2 SNPs contained *mucin 1* (*MUC1*) and other 4 genes and SNPs significantly associated with DGC (rs2070803: $P = 4.33 \times 10^{-13}$; odds ratio [OR], 1.71 by meta-analysis of the studies on the 3 panels) but not with intestinal-type gastric cancer. Functional studies demonstrated that rs4072037 ($P = 1.43 \times 10^{-11}$; OR, 1.66 by meta-analysis) in *MUC1* affects promoter activity and determines the major splicing variants of *MUC1* in the gastric epithelium. Individuals that carry both SNPs rs2294008 in *PSCA* and rs4072037 in *MUC1* have a high risk for developing DGC (OR, 8.38). **CONCLUSIONS: *MUC1* is the second major DGC susceptibility gene identified. The SNPs rs2070803 and rs4072037 in *MUC1* might be used to identify individuals at risk for this type of gastric cancer.**

Keywords: Stomach Cancer; Risk Genotype; Cancer Prevention; Genome-Wide Association Study.

Gastric cancer (GC) is the fourth most common cancer and the second most common cause of cancer death in the world.[1] More than 90% of GC are adenocarcinomas, which are classified into diffuse-type GC (DGC) and intestinal-type GC (IGC).[2] Typically, IGC arises through a sequence of pathologic changes of the gastric epithelium: chronic gastritis mainly because of *Helicobacter pylori* infection, atrophic gastritis, intestinal metaplasia, dysplasia, and adenocarcinoma.[3] On the other hand, the origin of DGC has been considered to be gastric epithelial stem cells and/or precursors present in the isthmus region of the middle portion of the epithelium (Supplementary Figure 1). Genetic and epigenetic events acting on the stem/precursor cells may cause a deviation from their normal differentiation program and lead to a DGC development,[4] although details are yet to be revealed. In contrast to the steady decline of the incidence of IGC, mainly because of the decreasing prevalence of *H pylori* infection, DGC appears to be increasing.[5] Moreover, some DGC develops to a highly malignant form, linitis plastica.[6] Identification of genetic predisposing factors and molecular pathways for the DGC development is one of the fundamentals for conceiving effective prevention, early diagnosis and therapeutic strategies.

Abbreviations used in this paper: DGC, diffuse-type gastric cancer; GC, gastric cancer; GWAS, genome-wide association study; IGC, intestinal-type gastric cancer; kb, kilobase; LD, linkage disequilibrium; *MUC1*, mucin 1; OR, odds ratio; por, poorly differentiated adenocarcinoma; *PSCA*, *prostate stem cell antigen*; sig, signet-ring cell carcinoma; TR, tandem repeats.

0016-5085/$36.00
doi:10.1053/j.gastro.2010.10.058

Previously, we conducted a gene-centric genome-wide association study (GWAS) on DGC and identified 3 statistically significant single nucleotide polymorphisms (SNPs) at 2 loci after Bonferroni correction ($P < 1.8 \times 10^{-5}$) in the second stage of the screening.[7] Of the 3 SNPs, rs2976392 showed the lowest *P* value and tagged a linkage disequilibrium (LD) block at chromosome 8q24.3, in which we identified *prostate stem cell antigen* (*PSCA*) encoding prostate stem cell antigen as the novel DGC susceptibility gene. In the present study, we investigated the second genomic region of interest at chromosome 1q22, which harbors the remaining 2 SNPs, rs2075570 and rs2070803,[7] and identified *mucin 1* (*MUC1*) as the possible causal gene of the association of the region to DGC. The association between the gene and GC had been suggested also in previous reports.[8–11] However, unlike the previous candidate gene approach, we have reached the gene by performing a hypothesis-free GWAS followed by biologic studies in which a rationale of the association was obtained through the analyses of the function of a SNP rs4072037. Moreover, this study has a sufficient power as a systematic survey of genetic factors with the expected range of effect size and allele frequencies, generating a convincing level of statistical association ($P < 10^{-10}$ as compared with $P \sim 10^{-2}$ by the previous candidate gene approach[11]). The SNP rs4072037 is known to determine a splicing acceptor site in the second exon of *MUC1*.[12] In this study, we showed that the SNP is also related to major splicing variant selection in the stomach and has effect on the *MUC1* promoter activity, both of which may result in MUC1 functional difference between the individuals.

Materials and Methods

Samples

In Japan, the common type of GC is classified into 7 categories: papillary adenocarcinoma (pap), tubular adenocarcinoma (tub1 and tub2), poorly differentiated adenocarcinoma (por1 and por2), signet-ring cell carcinoma (sig), and mucinous adenocarcinoma (muc). However, a classification into 2 major categories by Lauren,[2] intestinal and diffuse types, is used worldwide especially for clinicoepidemiologic studies. A review of the classification systems is described elsewhere.[7] Basically, the DGC under the Lauren classification corresponds to por2 (nonsolid type) of poorly differentiated adenocarcinoma and sig by Japanese classification, although some investigators consider that por1 (solid type) is also included in DGC.[13]

Details of DNA samples used in the SNP typing and the association study of the chromosome 1q22 locus are as follows: In the Tokyo data set study, 610 DNA samples from patients with DGC (320 males; mean age, 55.4; 290 females; mean age, 54.0) were prepared either from methanol-fixed, paraffin-embedded tissues of noncancerous gastric mucosa or lymph nodes, or from peripheral blood, of patients with either the linitis plastica type of GC or early-stage cancer diagnosed as macroscopic type 0 IIc with histologic type of por2 and/or sig. The DGC samples in the Tokyo data set were collected at 4 institutions: the National Cancer Center Hospital in Tokyo: 360 paraffin-embedded tissues and 164 blood samples; Nippon Medical School Hospital in Tokyo: 76 blood samples; Aichi Cancer Center in Aichi: 1 blood sample; and Shikoku Cancer Center in Ehime: 9 blood samples. The control DNA samples were from peripheral blood leukocytes of 1266 volunteer individuals (male, 849; mean age, 67.2; female, 417; mean age, 59.8) with no known malignancies, who offered blood at a health check examination at Iwata City Hospital in Shizuoka and at Keio University campuses in Tokyo.

In the Aichi data set study, the DGC case samples were obtained from peripheral blood of 304 patients with histologic diagnosis of por1, por2, or sig (199 males; mean age, 57.3; 105 females; mean age, 56.4). Control blood samples were from 1467 volunteer individuals (1098 males; mean age, 59.8; 369 females; mean age, 57.3) with no known malignancies. Power calculations for the DGC analysis showed that the sample size of 304 cases and 1467 controls would provide the study with a power of over 98% for detecting an association of a SNP with a minor allele frequency of 0.2 or higher and per-allele odds ratio (OR) for risk allele of 1.63 or higher (estimates on rs2070803 obtained from Tokyo data set) in a 2-sided test at a significance level of .05.

In the Korea data set study, peripheral blood samples were donated from 455 patients with DGC who were diagnosed or treated at the National Cancer Center in Seoul, Korea (260 males; mean age, 52.4; 195 females; mean age, 48.5). The control subjects were 372 volunteers who participated in the National Cancer Screening Program at the National Cancer Center, Korea, and were confirmed by endoscopy not to have GC (191 males; mean age, 54.2; 181 females; mean age, 52.5). Power calculations showed that the sample size of the Korea data set would provide the study with a power of over 95% for detecting an association of rs2070803 at a significance level of .05 for the DGC study.

In the association studies (results shown in Figure 1, Table 1, and Supplementary Tables 1–4), 11 subjects (4 DGC and 2 controls from Tokyo data set, 2 controls from Aichi data set, 3 DGC cases from Korea data set) were excluded because of at least 1 missing covariate. Distributions of the covariates from subjects included in the studies are shown in Supplementary Figures 9–11.

Haplotype-based association study was performed on DNA samples from 380 DGC cases (200 males; mean age,

BASIC-ALIMENTARY TRACT

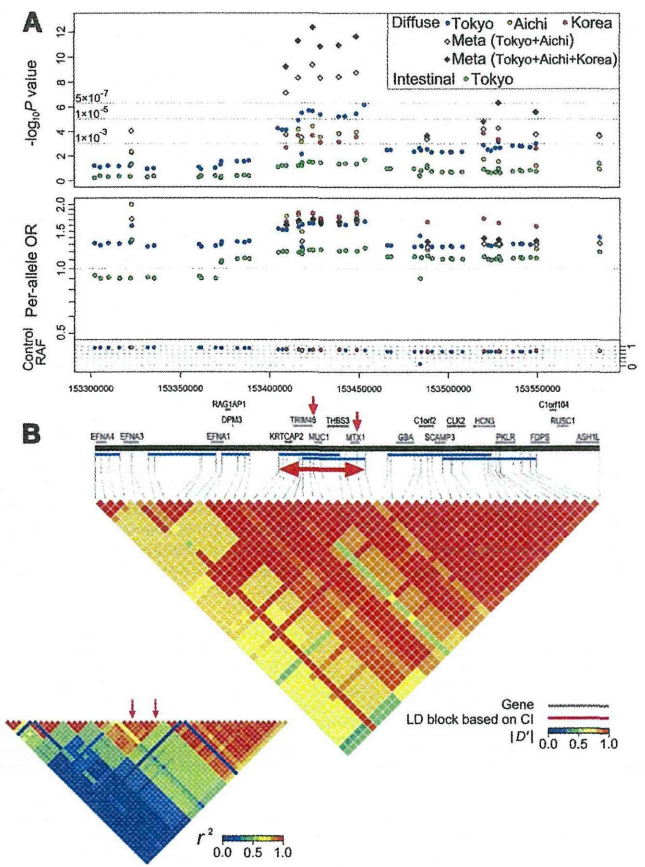

Figure 1. Association of the SNPs around rs2070803 and rs2075570 in chromosome 1q22 with GC and LD analyses of the SNPs. (*A*) The association study on DGC disclosed 8 SNPs with $P < 4 \times 10^{-5}$ within the LD block around rs2070803 and rs2075570 (*arrows* in *B*) in the Japanese population (Tokyo data set: *blue dots*). The association with DGC was replicated in another Japanese population (Aichi data set: *orange dots*) and also in the Korean population (Korea data set: *red dots*) for 4 selected SNPs: rs2070803, rs4072037, rs2066981, and rs2075570 ($P < 1 \times 10^{-3}$). Meta-analysis on these 3 data sets was also conducted (Tokyo and Aichi: *grey dots*; Tokyo, Aichi, and Korea: *black dots*). The study on IGC in the Japanese population (Tokyo data set: 599 cases, 1264 controls) showed no significant association of SNPs (*green dots*). *Upper panel* shows *P* value of each SNP in negative common logarithmic scale; *lower panel* shows OR and frequency of the risk allele (control RAF) of the SNPs. The position of the *dots* representing each SNP corresponds vertically to that in the physical map in *B*. (*B*) An LD analysis based on |*D'*| showed a strong LD around the 2 SNPs, rs2070803 and rs2075570 (*red arrows*), identified as DGC-associated SNPs in GWAS.[7] The strength of the LD is indicated by heat maps. *Blue bars* represent LD blocks defined by confidence intervals of |*D'*|.[15] Five genes—*KRTCAP2*, *TRIM46*, *MUC1*, *THBS3*, and *MTX1*—reside in the region of strong LD (*double-headed red arrow*) harboring rs2070803 and rs2075570. An LD map with r^2 is also shown in small scale. The analysis was performed with genotyping data of 1266 controls of the Tokyo data set.

Table 1. Results of Association Studies With 3 Independent Data Sets—Tokyo, Aichi, and Korea—for GC Susceptibility, on the 8 SNPs in the LD Block Containing rs2070803 and rs2075570 in Chromosome 1q22

SNPs, major/minor/risk[a]	Diffuse, Tokyo data set[b,c]	Diffuse, Aichi data set[c,d]	Diffuse, Korea data set[c,e]	Diffuse, Meta-analysis I[c,f]	Diffuse, Meta-analysis II[c,h]	Intestinal, Tokyo data set[c,i]
rs9426886 T/A/T	OR: 1.61 (1.32–1.97) *P*: 3.17×10^{-6} MAF (case): 0.133 MAF (control): 0.199					OR: 1.22 (1.01–1.46) *P*: 3.62×10^{-2} MAF (case): 0.169 MAF (control): 0.199
rs4971100 A/G/A	OR: 1.63 (1.33–1.99) *P*: 2.04×10^{-6} MAF (case): 0.133 MAF (control): 0.200					OR: 1.22 (1.02–1.47) *P*: 3.13×10^{-2} MAF (case): 0.169 MAF (control): 0.200
rs2070803[j] G/A/G	OR: 1.63 (1.33–1.99) *P*: 2.20×10^{-6} MAF (case): 0.133 MAF (control): 0.200	OR: 1.81 (1.36–2.40) *P*: 3.93×10^{-5} MAF (case): 0.104 MAF (control): 0.178	OR: 1.82 (1.32–2.49) *P*: 2.19×10^{-4} MAF (case): 0.103 MAF (control): 0.178	OR: 1.69 (1.43–1.99) *P*: 4.25×10^{-10}	OR: 1.71 (1.48–1.98) *P*: 4.33×10^{-13}	OR: 1.22 (1.02–1.47) *P*: 3.34×10^{-2} MAF (case): 0.168 MAF (control): 0.200
rs4072037 A/G/A	OR: 1.62 (1.32–1.99) *P*: 4.04×10^{-6} MAF (case): 0.126 MAF (control): 0.187	OR: 1.69 (1.27–2.25) *P*: 2.82×10^{-4} MAF (case): 0.099 MAF (control): 0.164	OR: 1.74 (1.26–2.39) *P*: 7.82×10^{-4} MAF (case): 0.093 MAF (control): 0.163	OR: 1.64 (1.39–1.94) *P*: 4.46×10^{-9}	OR: 1.66 (1.44–1.93) *P*: 1.43×10^{-11}	OR: 1.23 (1.02–1.48) *P*: 3.36×10^{-2} MAF (case): 0.157 MAF (control): 0.187
rs2066981 T/C/T	OR: 1.61 (1.31–1.98) *P*: 5.91×10^{-6} MAF (case): 0.125 MAF (control): 0.186	OR: 1.74 (1.31–2.32) *P*: 1.50×10^{-4} MAF (case): 0.099 MAF (control): 0.167	OR: 1.76 (1.27–2.43) *P*: 6.98×10^{-4} MAF (case): 0.092 MAF (control): 0.159	OR: 1.65 (1.40–1.95) *P*: 3.77×10^{-9}	OR: 1.67 (1.44–1.94) *P*: 1.11×10^{-11}	OR: 1.21 (1.01–1.46) *P*: 4.44×10^{-2} MAF (case): 0.158 MAF (control): 0.186
rs914615 G/A/G	OR: 1.61 (1.31–1.96) *P*: 5.83×10^{-6} MAF (case): 0.126 MAF (control): 0.187					OR: 1.22 (1.01–1.47) *P*: 3.94×10^{-2} MAF (case): 0.158 MAF (control): 0.187
rs2075570[j] A/G/A	OR: 1.63 (1.32–2.00) *P*: 3.45×10^{-6} MAF (case): 0.127 MAF (control): 0.189	OR: 1.77 (1.33–2.37) *P*: 1.12×10^{-4} MAF (case): 0.100 MAF (control): 0.170	OR: 1.84 (1.32–2.55) *P*: 2.73×10^{-4} MAF (case): 0.091 MAF (control): 0.163	OR: 1.67 (1.42–1.98) *P*: 1.73×10^{-9}	OR: 1.71 (1.47–1.98) *P*: 2.26×10^{-12}	OR: 1.21 (1.01–1.46) *P*: 4.25×10^{-2} MAF (case): 0.159 MAF (control): 0.189
rs1057941 T/C/T	OR: 1.67 (1.36–2.04) *P*: 6.59×10^{-7} MAF (case): 0.130 MAF (control): 0.200					OR: 1.25 (1.04–1.50) *P*: 1.95×10^{-2} MAF (case): 0.168 MAF (control): 0.200

NOTE. OR: per-allele odds ratio for risk allele with 95% confidence intervals in parentheses.
MAF, minor allele frequency; SNP, single nucleotide polymorphism.
[a]Major, minor, and risk alleles are common in Tokyo, Aichi, and Korea data sets.
[b]Association study on diffuse-type GC in the Japanese population (Tokyo data set, 606 cases, 1264 controls), performed with fine-mapping data.
[c]*P* values and ORs with 95% CI were calculated under an additive model using logistic regression adjusted for age, gender, and number of risk alleles of rs2294008 of the *PSCA* gene, which was associated with DGC in our previous GWAS (Study Group of Millennium Genome Project for Cancer[7]).
[d]Association study on diffuse-type GC in the Japanese population (Aichi data set: 304 cases, 1465 controls), for replication of the study on Tokyo data set.
[e]Association study on diffuse-type GC in the Korean population (Korea data set: 452 cases, 372 controls), for replication of the study on Tokyo data set.
[f]Meta-analysis on the data of the Tokyo and Aichi data sets.
[g]*P* values and ORs with 95% CI were calculated using a random effects model.
[h]Meta-analysis on the data of the Tokyo, Aichi, and Korea data sets.
[i]Association study on intestinal-type GC in the Japanese population (Tokyo data set: 599 cases, 1264 controls).
[j]SNPs identified in the GWAS (Study Group of Millennium Genome Project for Cancer[7]).

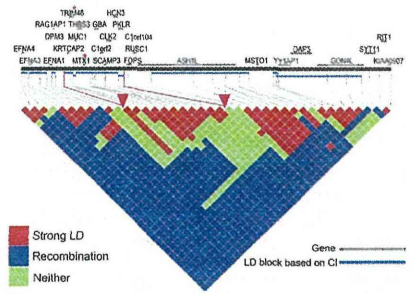

Figure 2. LD analyses on the SNPs in chromosome 1q22 using Gabriel's criteria[15] based on genotyping data of 680 Japanese controls. LD blocks with the criteria are in *blue horizontal lines*. *Red asterisks* indicate positions of rs2075570 and rs2070803. The LD block (both ends indicated by *red triangles*), in which the 2 SNPs reside, contains 12 SNPs and 11 genes. CI, confidence interval.

56.1; 178 females; mean age, 54.6; 2 missing gender information; 372 of which were the same as DGC cases in the Tokyo data set) and 309 controls (151 males; mean age, 49.1; 158 females; mean age, 46.1; 306 of which were included in the Tokyo data set). DGC samples were collected at 2 institutions as follows: 318 paraffin-embedded tissues at the National Cancer Center Hospital, and 62 blood samples at Nippon Medical School Hospital. Control DNA samples were from Keio University campuses.

The Japanese part of the study was approved by the ethics committees of the participating institutions in accordance with the Ethics Guidelines For Human Genome/Gene Analysis Research in Japan. The Korean side of the GC case control study was approved by the Ethics Committee of the National Cancer Center, Korea. Informed consent was obtained from all living subjects, including opt-out consent for the paraffin block archival samples.

LD Analysis

The LD map of chromosome 1q22 (Figure 2) was constructed based on the genotype data of 41 SNPs (Supplementary Table 10) obtained from 680 Japanese controls (436 males; mean age, 43.7; 242 females; mean age, 43.7; 2 missing gender information; 371 of which were from controls in the Tokyo data set) genotyped by Illumina Human610-Quad BeadChip (Illumina, San Diego, CA). The LD map shown in Figure 1 was constructed based on the fine mapping data of the 52 SNPs (Supplementary Table 10) from the Single Nucleotide Polymorphism database (http://www.ncbi.nlm.nih.gov/projects/SNP/) on the 1266 Japanese controls (the same as controls in the Tokyo data set). The pattern of LD was analyzed using 2 parameters, r^2 and $|D'|$,[14] and the confidence interval of the $|D'|$ was also utilized.[15]

Statistical Analyses

Statistical significance of the association was evaluated for each SNP by logistic regression. *P* values under an additive model adjusted for 3 age categories (≤39, 40–59, and ≥60 years), gender, and the risk alleles at rs2294008 in *PSCA* (Supplementary Figures 9–11). The significance level was set to .05 by Bonferroni correction for multiple testing, meaning $P = 8.9 \times 10^{-4}$ before correction for the Tokyo data set. Meta-analyses of the Tokyo data set and the Aichi and Korea replication data sets were performed using a random effects model.[16] Haplotype-based association was tested by Fisher exact test. Haplotype phases in each individual were inferred by fastPHASE software.[17] Other statistical analyses were carried out using the R suite (http://www.r-project.org/) and the StatXact 8 (Cytel Inc, Cambridge, MA). Population stratification of the Tokyo data set was examined previously by the STRUCTURE software,[18] the Genomic Control, and mixture model methods,[19,20] and no significant subpopulation was detected.[7]

In the association studies using 2-locus genotype data of rs4072037 and rs2294008, the biologic effect of the SNPs' risk allele was assumed to be recessive (rs4072037) or dominant (rs2294008), ie, the risk genotype for rs4072037 is AA and, for rs2294008, TT and TC (Figure 3). Risk factor variables consist of 4 categories based on the genotypes of rs4072037 and rs2294008. *P* value and OR and its 95% confidence interval (CI) for each category was obtained by logistic regression adjusted for age and gender.

Other Analyses

The materials and methods used in IGC association studies, genotyping, resequencing, and functional studies are described in Supplementary Materials and Methods and Supplementary Tables 9–11.

Results

Identification of the Susceptibility Region in Chromosome 1q22

Initially, we analyzed LD ($|D'|$) around the 2 marker SNPs based on the genotyping data on 680 control subjects. The criteria based on a confidence interval of $|D'|$[15] was applied to find an LD block containing 12 SNPs (including rs2075570 and rs2070803) and 11 genes (Figure 2). The second analysis of high-density genotyping around this block was performed on 610 cases of DGC and 1266 controls (Tokyo data set) for 52 SNPs selected from the Single Nucleotide Polymorphism database. A solid 49kb-LD block was identified spanning 13 SNPs including rs2075570 and rs2070803. Eight SNPs in the block showed strong associations ($P < 1.0 \times 10^{-5}$)

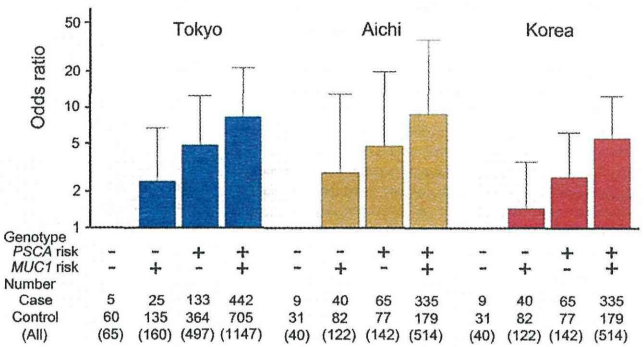

Figure 3. Association studies for DGC using 2-locus genotype data of rs4072037 in *MUC1* and rs2294008 in *PSCA*. The association studies were performed with a distinct model for each risk allele's effect, recessive for rs4072037 and dominant for rs2294008. *Bar*, upper bound of 95% confidence interval.

with DGC (Figure 1 and Table 1). Of the 8 SNPs, 4 were selected and genotyped on independent case-control sets in Japan (Aichi data set: 304 cases, 1467 controls) and in Korea (Korea data set: 455 cases, 372 controls), and the association was replicated in both data sets (Figure 1 and Table 1). A meta-analysis of the 3 case-control studies also showed significant correlation of the region: $P = 2.26 \times 10^{-12}$; OR, 1.71 for rs2075570 and $P = 4.33 \times 10^{-13}$; OR, 1.71 for rs2070803. Because this 1q22 region was originally identified by the GWAS on DGC,[7] an association with IGC was examined on the 3 populations: 601, 274, and 415 cases from Tokyo, Aichi, and Korea, respectively, using the same control subjects analyzed for DGC. The 1q22 region was not significantly associated with IGC (Figure 1, Supplementary Figure 2). Full data of the association studies are shown in Supplementary Tables 1–4.

Polymorphisms in MUC1 *Gene and Haplotype-Based Association Study*

The 49-kilobase (kb) block contained 5 genes encoding keratinocyte associated protein 2 (*KRTCAP2*), tripartile motif protein 46 (*TRIM46*), mucin 1 (*MUC1*), thrombospondin 3 (*THBS3*), and metaxin 1 (*MTX1*). Based on their expression patterns and gene annotations, we prioritized *MUC1* for further analyses because *MUC1* is expressed in pit cells in the pit region, mucous neck cells in the neck region, chief (zymogenic) cells in the base region, and parietal cells in the neck and base regions of the gastric epithelium (Figure 4*A*, Supplementary Figure 1).[21] Moreover, previous studies based on a candidate-gene approach reported an association between its polymorphisms and GC.[8–11]

The resequencing of the *MUC1* gene identified a total of 7 polymorphisms in 48 Japanese individuals: 4 SNPs without rs numbers (numbers 1, 2, 4, and 7), 1 indel (No. 6, rs66597679), rs12411216 (No. 3), and rs4072037 (No. 5) (Figure 4*B* and Supplementary Table 5). The 7 were genotyped on 380 Japanese cases and 309 controls (Table 2), and, in a subsequent haplotype analysis, SNPs numbers 1 and 7 were removed from analysis because they were monomorphic in the 689 Japanese individuals. The remaining 5 SNPs were used for a haplotype-based association study, which revealed 3 major haplotypes, numbers 1–3, with ORs of 1.32, 0.90, and 0.65, respectively, and 1 minor haplotype, No. 4, with minor allele frequency of 0.0105 in cases and 0.0032 in controls (Table 3).

Functional Analyses of MUC1 *SNPs*

Seven transcriptional variants are registered as *MUC1* messenger RNA in the National Center for Biotechnology Information database (http://www.ncbi.nlm.nih.gov/) (Supplementary Figure 3), and the rs4072037 SNP ($P = 1.43 \times 10^{-11}$ and OR of 1.66 by meta-analysis of the 2 Japanese and 1 Korea data sets, Table 1) located in exon 2 of *MUC1* had been found to be related to the splicing site selection in the exon.[12] To identify the variants expressed in the stomach, we conducted RNA ligase-mediated rapid amplification of the 5′ complementary DNA end procedure.

Our results showed that the major transcripts in the stomach are variants 2 and 3 (Figure 4*B* and Supplementary Figure 4) and that all the examined clones of variant 2 possessed G allele at rs4072037, in contrast to those of the variant 3 possessing the A allele, as reported previously (Supplementary Figure 5).[12] This suggests that rs4072037 is significantly involved in the splicing regulation of the second exon. In other words, it is likely that the SNP directly determines the relative dominance of the 2 major *MUC1* splicing variants, the variants 2 and 3, in the gastric epithelium.

As reported previously on the Caucasian population,[12] no polymorphisms other than rs4072037 were found in the region spanning from exon 1 to 2, which might affect the splicing of the second exon, by our resequencing of the

BASIC-ALIMENTARY TRACT

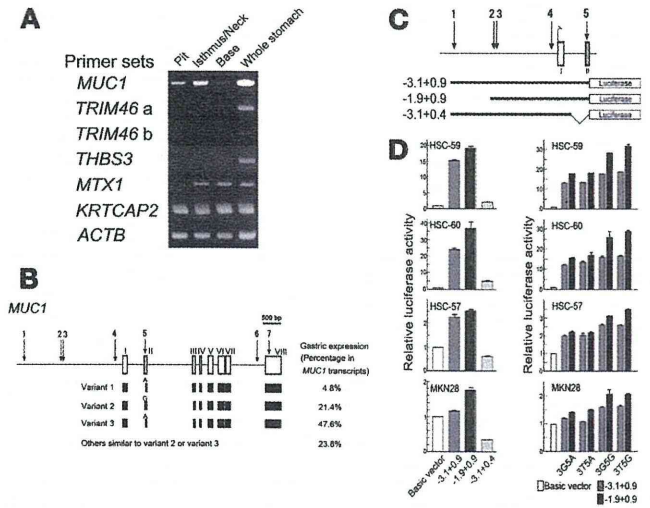

Figure 4. Functional analyses of MUC1 and its SNPs. (*A*) Expression analysis on the 5 genes in the LD block associated with DGC using microdissected gastric samples (reverse-transcription polymerase chain reaction). (*B*) Seven polymorphisms in the *MUC1* gene identified by resequencing of 48 Japanese controls and *MUC1* transcriptional variants detected in RNAs from the gastric mucosa by RNA ligase-mediated rapid amplification of the 5′ complementary DNA (cDNA) end procedure (5′RACE). SNP positions are indicated by *arrows* with numbers corresponding to Table 2 and Supplementary Table 5. The 5′RACE was conducted on a pooled gastric RNA sample from 21 Caucasian individuals. All the variant 2 clones contained exclusively the G allele of SNP No. 5 and all the variants 1 and 3 the A allele, without exception. Complete result of the 5′RACE is presented in Supplementary Figure 4. (*C*) *MUC1* genomic fragments examined in reporter assays. (*D*) Effect of SNP No.5 (rs4072037) on the transcriptional activity of the *MUC1* promoter (reporter assay). The transcriptional activity of the −3.1 to +0.9 region was significantly reduced by truncating the +0.4 to +0.9 region (*−3.1+0.4*) in gastric cancer cell lines HSC-59, HSC-60, HSC-57, and MKN28. The genomic fragments with G allele in the SNP (*3G5G* and *3T5G*) showed higher reporter activity than that with A allele (*3G5A* and *3T5A*).

MUC1 gene on 48 Japanese individuals (Figure 4*B* and Supplementary Table 5). As regards the splicing variation, the risk allele A at rs4072037, which is found in variants 1 and 3 (Supplementary Figure 5), causes a 9-amino acid deletion in the second exon and consequently modifies both the signal peptide and N-terminal amino acid of the mature protein by changing the signal-peptide cleavage site.[12] This may change the intracellular trafficking and glycosylation and folding of the protein, leading to alteration in the function of the mature protein.

To examine the difference in the MUC1 function among the variants, we evaluated in vitro functions of the

Table 2. Seven SNPs and Their Association With DGC Based on the Genotype Data of 380 Japanese Cases and 309 Controls

SNP no.	rs number	Major allele	Minor allele	Risk allele	MAF (case)	MAF (control)	OR[a]	95% CI	*P* value[b]
1		T	C		0.0000	0.0000			
2		C	T	C	0.1882	0.2039	1.11	0.85–1.44	.4945
3	12411216	G	T	G	0.1289	0.1851	1.53	1.14–2.06	.004344
4		G	A	A	0.0105	0.0032	3.28	0.69–15.49	.1998
5	4072037	A	G	A	0.1289	0.1857	1.54	1.15–2.07	.004275
6	66597679	AC	—	AC	0.1308	0.1869	1.53	1.14–2.05	.005225
7		C	T		0.0000	0.0000			

CI, confidence interval; MAF, minor allele frequency.
[a]Odds ratio for risk allele.
[b]*P* values obtained by Fisher exact test.

Table 3. Four Major Haplotypes Inferred and Their Association With DGC Based on the Genotype Data of 380 Japanese Cases and 309 Controls

	SNP no.									*P* values	
	2	3	4	5	6	Case[a]	Control[b]	OR	95% CI	Fisher[c]	Permutation[d]
Haplotype no.											
1	C	G	G	A	AC	0.6733	0.6084	1.32	1.06–1.65	.01501	.0483
2	T	G	G	A	AC	0.1884	0.2039	0.90	0.69–1.18	.4945	.9027
3	C	T	G	G	—	0.1278	0.1845	0.65	0.48–0.87	.004200	.0127
4	C	G	A	A	AC	0.0105	0.0032	3.28	0.69–15.49	.1998	.4058

CI, confidence interval.
[a]Frequency of case.
[b]Frequency of control.
[c]*P* values obtained by Fisher exact test.
[d]*P* values obtained by permutation test (100,000 permutations performed).

2 major variants expressed in the gastric epithelium: variants 2 and 3. Because the full-length product of *MUC1* is well-known for its cell growth-promoting activity in cancer cells,[22,23] we individually transfected a cytomegalovirus promoter-driven complementary DNA of *MUC1* variant 2 or 3 to the MKN28 cells, which express *MUC1* at an undetectable level (data not shown). Examination of their cell growth by both cell counting and colorimetric methods suggested that variant 2 is more potent in growth-promoting activity than variant 3 (Supplementary Figures 6 and 7). Although the observed difference seems to be small, this level of normal range of individual variation is generally expected for a common genetic variant influencing a common disease susceptibility and is probably because of the difference in the signal peptide or the N-terminal structure of the mature protein encoded by each variant because the other portion of the amino acid sequence is common between the 2 variants.

Next, we investigated the function of the *MUC1* SNPs in the context of the haplotypes. We selected haplotype No. 1 as the major risk haplotype and haplotype No. 3 as the most protective haplotype and analyzed the functions of SNPs numbers 3 and 5, excluding SNP No. 6 from our functional analyses because its location in the intron 7 made it unlikely to be involved in the transcriptional regulation and/or alternative splicing of the gene. SNP numbers 2 and 4 were also excluded because they were found on both the risk and the protective haplotypes. Because the remaining 2 SNPs, numbers 3 and 5, do not change amino acid, we first examined, by a reporter assay, their effect on the transcriptional regulation of *MUC1*; the region spanning −1.9 to 0.9 kb relative to the *MUC1* transcription start site had a transcriptional activity (Figure 4*C* and *D*). The reporter assay on base-substituted constructs showed that, in all the gastric carcinoma cell lines examined, the fragments containing the G allele at rs4072037 (SNP No. 5), which is present only in the protective haplotype, has a higher transcriptional activity than that with an A allele present in the risk haplotype (Figure 4*D*, *right panel*). The assay on truncated constructs showed that a removal of a +0.4 to +0.9-kb region, which contains rs4072037, significantly diminishes the transcriptional activity (Figure 4*D*, *left panel*). We also confirmed in the reporter assay that the T allele of SNP No. 2, which is unique to haplotype No. 2 showing OR of 0.9 but no significant *P* value, has no effect (Supplementary Figure 8).

In sum, the results in this study and from previous reports by other investigators suggest that rs4072037 SNP has at least 2 functions: (1) regulation of the alternative splicing at the second exon and (2) modification of the transcriptional activity of the promoter. The association study in the context of LD and the functional study strongly implicate rs4072037 as a functional cause of the association between the 1q22 region and DGC susceptibility.

Association Studies for DGC Using 2-Locus Genotype Data of rs4072037 in MUC1 *and rs2294008 in* PSCA

Finally, we examined the effect of 2 DGC susceptibility SNPs identified by our GWAS, rs4072037 in *MUC1* and rs2294008 in *PSCA*, both of which are functional, using Tokyo, Aichi, and Korea data sets. When a genetic model is tentatively selected for each locus by simply comparing *P* value, a recessive and dominant model was applied for rs4072037 and for rs2294008, respectively; the individuals possessing the risk genotype of both SNPs showed significant risk for developing DGC (eg, OR, 8.38 in Tokyo data set, Figure 3). Notably, individuals with protective alleles of both SNPs were observed only in controls (Supplementary Table 6).

Discussion

At chromosome 1q22, we focused on the region with strong LD around rs2075570 and rs2070803 using Gabriel et al's criteria.[15] The region contains 5 genes: *TRIM46*, *THBS3*, *MTX1*, *KRTCAP2*, and *MUC1* (Figure 1). We prioritized the genes for the subject of

further studies by the first criterion (whether the gene is expressed in the gastric epithelial cells) and then by the second one (whether annotated function suggests its involvement in carcinogenesis). We observed transcripts of 4 of the 5 genes in microdissected samples of the gastric epithelium by reverse-transcription polymerase chain reaction, but no transcript of *TRIM46* was detectable there (Figure 4*A*). *THBS3* encoding a multifunctional extracellular matrix glycoprotein is expressed in multiple human tissues including the stomach,[24] and no evidence of a causal relation to carcinogenesis has been obtained. *MTX1* encodes a component of a preprotein import complex in the outer membrane of the mammalian mitochondrion.[25] If it is involved in carcinogenesis, the effect of its SNP would be reflected in many types of cancer, yet no such involvement has surfaced. *KRTCAP2* encodes a protein possessing transmembrane domain, showing multitissue expression.[26] Its function is unknown, and its relation to carcinogenesis has not been demonstrated.

In this study, we considered *MUC1* as a strong candidate for the gene responsible for the association of 1q22 with DGC because, in addition to several previous candidate gene analyses showing an association with *MUC1* polymorphisms and GC,[8–11] MUC1 has been considered to possess an oncogenic property as described below.

Mucin family members are classified into 2 types, secreted or membranous, based on their localization, and MUC1 is a transmembrane mucin.[27] MUC1 is a multifunctional protein involved in mucosal lubrication, protection from pathogens, signal transduction, and cell-cell interaction.[27] MUC1 was over-expressed in breast, ovarian, lung, pancreatic, and prostate cancers and was a marker of poor prognosis in gastric cancer.[21,28] Several in vivo studies have provided evidence supporting its function in carcinogenesis. MUC1 has a role in cell growth, anchorage independence, cell migration, antiapoptotic property, and drug resistance of cancer cells,[22,29–33] all of which are accomplished through interaction with several signaling pathways,[34] although these lines of biologic evidence were obtained on the standard molecule containing tandem repeats (TR). Because the MUC1 expressed in the gastric epithelium has no TR, it is possible that its function in gastric epithelial cells is different from that of the TR-containing product in other epithelial cells.[35–37]

Recently, however, it is supposed that MUC1 has a protective function against environmental insults and acts against tumorigenesis in normal epithelial cells, which keep maintaining their cell polarity. In contrast, once the cells lose cell polarity in consequence of prolonged inflammation, MUC1 promotes cell growth and acts for tumorigenesis.[38] It was also reported that MUC1 functions as a growth factor receptor in human embryonic stem cells.[39] It is presumable that MUC1 is involved in growth regulation of gastric stem cells and progenitors, which are considered to be the origin of DGC. Like the function of PSCA, which is up-regulated in prostate and urinary bladder cancers but suppressed in gastric cancer,[7] the MUC1 function may differ between cell types, normal or malignant, and among different tissues. In the same manner as PSCA, MUC1 is down-regulated in intestinal metaplasia of the gastric epithelium from which IGC arises.[40] If MUC1 has some protective function in carcinogenesis, this down-regulation makes the stem and progenitor cells more susceptible to carcinogenic events. In any case, further research is needed to explore the pleiotropic functions of *MUC1*.

In this study, we demonstrated that rs4072037 has a role in transcriptional regulation and also in splicing site selection leading to the dominant variant determination of *MUC1* transcripts in gastric epithelial cells. If variant 3 is less functional in protection against DGC than variant 2, the possession of the A allele in the genome confers both quantitatively and qualitatively unfavorable consequences to MUC1 function, which may result in additive risk for DGC susceptibility.

In our GWAS on DGC susceptibility, the 2 loci showing the highest statistical significance directed us to the 2 functional SNPs: rs4072037 in *MUC1* and rs2294008 in *PSCA*.[7] It is noteworthy that both SNPs in the 2 genes appear to have dual functions: transcriptional regulation and signal-peptide modification.[7] Further investigation is required to validate the role of *MUC1* in DGC susceptibility and details of the mechanism that links the risk haplotype tagged by rs4072037 to DGC development.

The risk allele A of rs4072037 is in strong LD with the small allele of the variable numbers of tandem repeats in the second intron of the *MUC1* both in Europeans and Japanese; more than 90% of chromosomes have a nonrecombinant haplotype in both populations (Ng et al[12] and Supplementary Table 7), and the small allele was associated with GC in the European population.[8,9,12] However, the variable numbers of tandem repeats is unlikely to be the causal polymorphism for DGC susceptibility because the TRs are translated neither in normal nor malignant gastric epithelial cells (Supplementary Figure 4 and Supplementary Table 8).

This study has not only replicated the association of the *MUC1* SNP with GC in the Japanese and Korean populations, in addition to the previous reports on the Chinese and Caucasian population GCs,[8–11] but it has also disclosed that the association appears specific to DGC. Following discovery of the DGC-specific association of the *PSCA* polymorphism, this study has offered another piece of evidence to support distinct mechanisms for DGC and IGC development.

Although there was no significant interaction between the *MUC1* and *PSCA* SNPs for the DGC risk in our study (Figure 3, Supplementary Table 6, and data not shown), it is estimated that individuals with the double risk genotype are the majority in Japanese (56%) and Korean (49%) populations with a significant OR, 8.38, in Japanese, as compared with the lowest risk category. GWAS and other emerging genome analysis tools may unveil a number of polymorphisms showing a significant statistical association, but it is important to identify functional SNPs potentially related to carcinogenesis. The accumulation of information on the functional SNPs, environmental factors, and their interactions, all of which are truly related to DGC susceptibility, will make the genotyping a more practical tool for evaluating the individual risk for DGC and offer effective prevention strategies in the future.

Supplementary Material

Note: To access the supplementary material accompanying this article, visit the online version of *Gastroenterology* at www.gastrojournal.org, and at doi: 10.1053/j.gastro.2010.10.058.

References

1. Brenner H, Rothenbacher D, Arndt V. Epidemiology of stomach cancer. In: Verma M, ed. Methods of molecular biology: cancer epidemiology. Volume 472. New Jersey: Humana, 2009:467–477.
2. Lauren P. The two histological main types of gastric carcinoma: diffuse and so-called intestinal-type carcinoma. An attempt at a histo-clinical classification. Acta Pathol Microbiol Scand 1965; 64:31–49.
3. Chiba T, Marusawa H, Seno H, et al. Mechanism for gastric cancer development by *Helicobacter pylori* infection. J Gastroenterol Hepatol 2008;23:1175–1181.
4. Schier S, Wright NA. Stem cell relationships and the origin of gastrointestinal cancer. Oncology 2005;69(Suppl 1):9–13.
5. Henson DE, Dittus C, Younes M, et al. Differential trend in the intestinal and diffuse types of gastric carcinoma in the United States, 1973–2000—increase in the signet ring cell type. Arch Pathol Lab Med 2004;128:765–770.
6. Rosai J. Rosai and Ackerman's surgical pathology. Edinburgh: Mosby, 2004.
7. Study Group of Millennium Genome Project for Cancer. Genetic variation in *PSCA* is associated with susceptibility to diffuse-type gastric cancer. Nat Genet 2008;40:730–740.
8. Carvalho F, Seruca R, David L, et al. *MUC1* gene polymorphism and gastric cancer—an epidemiological study. Glycoconj J 1997; 14:107–111.
9. Silva F, Carvalho F, Peixoto A, et al. *MUC1* gene polymorphism in the gastric carcinogenesis pathway. Eur J Hum Genet 2001;9: 548–552.
10. Xu Q, Yuan Y, Sun LP, et al. Risk of gastric cancer is associated with the *MUC1* 568 A/G polymorphism. Int J Oncol 2009;35: 1313–1320.
11. Jia Y, Persson C, Hou L, et al. A comprehensive analysis of common genetic variation in *MUC1*, *MUC5AC*, *MUC6* genes and risk of stomach cancer. Cancer Causes Control 2010;21:313–321.
12. Ng W, Loh AX, Teixeira AS, et al. Genetic regulation of *MUC1* alternative splicing in human tissues. Br J Cancer 2008;99:978–985.
13. Japanese Gastric Cancer Association. Japanese classification of gastric carcinoma. 2nd English edition. Gastric Cancer 1998;1: 10–24.
14. Nei M. Molecular evolutionary genetics. New York: Columbia University Press, 1987.
15. Gabriel SB, Schaffner SF, Nguyen H, et al. The structure of haplotype blocks in human genome. Science 2002;296:2225–2229.
16. DerSimonian R, Laird N. Meta-analysis in clinical trials. Control Clin Trials 1986;7:177–188.
17. Scheet P, Stephens M. A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. Am J Hum Genet 2006; 78:629–644.
18. Falush D, Stephens M, Pritchard JK. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. Genetics 2003;164:1567–1587.
19. Bacanu SA, Devlin B, Roeder K. The power of genomic control. Am J Hum Genet 2000;66:1933–1944.
20. Zhu X, Zhang S, Zhao H, et al. Association mapping, using a mixture model for complex traits. Genet Epidemiol 2002;23: 181–196.
21. Senapati S, Sharma P, Bafna S, et al. The MUC gene family: their role in the diagnosis and prognosis of gastric cancer. Histol Histopathol 2008;23:1541–1552.
22. Li Y, Liu D, Chen D, et al. Human DF3/MUC1 carcinoma-associated protein functions as an oncogene. Oncogene 2003;4: 6107–6110.
23. Tsutsumida H, Swanson BJ, Singh PK, et al. RNA interference suppression of MUC1 reduces the growth rate and metastatic phenotype of human pancreatic cancer cells. Clin Cancer Res 2006;15:2976–2987.
24. Adolph KW. Relative abundance of Thrombospondin 2 and Thrombospondin 3 mRNAs in human tissues. Biochem Biophys Res Commun 1999;258:792–796.
25. Armstrong LC, Saenz AJ, Bornstein P. Metaxin 1 interacts with metaxin 2, a novel related protein associated with the mammalian mitochondrial outer membrane. J Cell Biochem 1999;74: 11–22.
26. Bonkobara M, Das A, Takao J, et al. Identification of novel genes for secreted and membrane-anchored proteins in human keratinocytes. Br J Dermatol 2003;148:654–664.
27. Gendler SJ. MUC1, the renaissance molecule. J Mammary Gland Biol Neoplasia 2001;6:339–353.
28. Taylor-Papadimitriou J, Burchell J, Miles DW, et al. MUC1 and cancer. Biochim Biophys Acta 1999;1455:301–313.
29. Suwa T, Hinoda Y, Makiguchi Y, et al. Increased invasiveness of *MUC1* cDNA-transfected human gastric cancer MKN74 cells. Int J Cancer 1998;76:377–382.
30. Raina D, Kharbanda S, Kufe D. The MUC1 oncoprotein activates the anti-apoptotic phosphoinositide 3-kinase/Akt and Bcl-xL pathways in rat 3Y1 fibroblasts. J Biol Chemi 2004;279:20607–20612.
31. Ren J, Agata N, Chen D, et al. Human MUC1 carcinoma-associated protein confers resistance to genotoxic anticancer agents. Cancer Cell 2004;5:163–175.
32. Huang L, Chen D, Liu D, et al. MUC1 oncoprotein blocks glycogen synthase kinase 3β-mediated phosphorylation and degradation of β-catenin. Cancer Res 2005;65:10413–10422.
33. Wei X, Xu H, Kufe D. Human mucin 1 oncoprotein represses transcription of the *p53* tumor suppressor gene. Cancer Res 2007;67:1853–1858.
34. Singh PK, Hollingsworth MA. Cell surface-associated mucins in signal transduction. Trends Cell Biol 2006;16:467–476.
35. Gendler SJ, Lancaster CA, Taylor-Papadimitriou J, et al. Molecular cloning and expression of human tumor-associated polymorphic epithelial mucin. J Biol Chem 1990;265:15286–15293.

36. Zrihan-Licht S, Vos HL, Baruch A, et al. Characterization and molecular cloning of a novel MUC1 protein, devoid of tandem repeats, expressed in human breast cancer tissue. Eur J Biochem 1994;224:787–795.
37. Baruch A, Hartmann M, Zrihan-Licht S, et al. Preferential expression of novel MUC1 tumor antigen isoforms in human epithelial tumors and their tumor-potentiating function. Int J Cancer 1997; 71:741–749.
38. Kufe DW. Mucins in cancer: function, prognosis and therapy. Nat Rev Cancer 2009;9:874–885.
39. Hikita ST, Kosik KS, Clegg DO, et al. MUC1* mediates the growth of human pluripotent stem cells. PLoS One 2008;3:e3312.
40. Silva E, Teixeira A, David L, et al. Mucins as key molecules for the classification of intestinal metaplasia of the stomach. Virchows Arch 2002;440:311–317.

Received May 8, 2010. Accepted October 26, 2010.

Reprint requests
Address requests for reprints to: 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; fax: (81) 3-3541-2685.

Acknowledgments
The authors thank Sachiyo Mimaki, Mineko Ushiama, Chie Naito, Yoko Odaka, Misuzu Okuyama, Miki Watanabe, and Hidemi Ito for their technical contributions and Seiji Ito for the clinical data collection of the Aichi data set.

Conflicts of interest
The authors disclose no conflicts.

Funding
Supported in Japan by the program for promotion of Fundamental Studies in Health Sciences of the National Institute of Biomedical Innovation (NiBio); in Aichi, supported by a Grant-in-Aid for Scientific Research from the Ministry of Education, Science, Sports, Culture and Technology of Japan and by a Grant-in-Aid for the Third Term Comprehensive 10-Year Strategy for Cancer Control from the Ministry of Health, Labour and Welfare of Japan; and, in Korea, part of the study was supported by grant 0710340 from the National Cancer Center, Korea.

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