Table 2. Clinical Factors Associated With Expression of Hepatic Interferon-Stimulated Genes

	Univariate analysis				Multivariate analysis			
Clinical factor	β	95% CI		P value	β	95% CI		P value
AST (*IU/L*)	0.274	0.13	0.42	<.001	—	—	—	
γ-GTP (*IU/L*)	0.326	0.18	0.47	<.001	0.288	0.14	0.43	<.001
HCV-RNA (*KIU/mL*)	−0.170	−3.19	−0.02	.025	−0.255	−0.40	−0.11	<.001
SVR+TR	−0.237	−0.32	−0.05	.009	—	—	—	
NR	−0.168	−0.57	0.18	.298	—	—	—	
FBS (*mg/dL*)	0.182	0.03	0.35	.021	—	—	—	
Insulin (*μU/mL*)	0.190	0.03	0.34	.016	—	—	—	
HOMA-IR	0.181	0.03	0.33	.017	—	—	—	.073
TG (*mg/dL*)	0.201	0.05	0.35	.011	—	—	—	.089
LDL-Chol (*mg/dL*)	−0.177	−0.33	−0.02	.025	−0.143	−0.28	0.00	.048

γ-GTP, γ-glutamyl transpeptidase; AST, aspartate aminotransferase; FBS, fasting blood sugar; TG, triglycerides; TR, transient response; NR, no response; SVR, sustained viral response; HOMA-IR, homeostasis model assessment of insulin resistance; LDL-chol, low-density lipoprotein cholesterol; CI, confidence interval; β, β coefficient; CI, confidence interval.

Expression of Hepatic ISGs Before Treatment Is Associated With Genetic Variation of IL28B

Recently, a GWAS successfully identified the genomic locus associated with the treatment response to Peg-IFN and RVB combination therapy for CH-C. Genetic variation in IL28B predicts HCV treatment-induced viral clearance.[11,12] We determined the genetic variation in IL28B of 32 patients[12] (Table 3). The SNPs rs8105790, rs11881222, rs8099917, and rs7248668 had a significant association with treatment response (odds ratio: 24.7–27.1, $P = 1.84 \times 10^{-30}$–2.68×10^{-32}). These SNPs are located in block 2 of the IL28B haplotype and show significant linkage disequilibrium in the HapMap data.[12] Ge et al[11] reported a different SNP (rs12979860) that was located between rs11881222 and rs8099917. The nucleotide sequence of rs12979860 was determined by direct sequencing, and the results are shown in Table 3. There was a strong association of rs12979860 and the other 4 SNPs indicating that this SNP was located within the same haplotype block. We confirmed these findings in multiple samples from Japanese patients (data not shown).

We selected rs8099917 for further study and evaluated it using TaqMan Pre-Designed SNP Genotyping Assays. The G nucleotide of rs8099917 was associated with a poor response to treatment (minor allele), whereas the T was associated with a fair response to treatment (major allele).[12] Out of 91 patients (Supplementary Table 3), the proportion of major homozygotes (TT), heterozygotes (TG), and minor homozygotes (GG) were 66% (60/91), 30% (27/91), and 4% (4/91), respectively (Table 4); 86% (51/60) of the major genotype (TT) patients had SVR or TR, whereas 65% (20/31) with the minor genotypes (TG or GG) had NR ($P < .001$).

Interestingly, hepatic gene expression profiles revealed that patients with the minor genotype showed higher expression of hepatic ISGs, whereas patients with the major genotype showed lower expression of hepatic ISGs (Figures 2 and 3). To examine further the relationship of the genetic variation in IL28B and its expression levels, we evaluated the expression of IL28B in the liver by RTD-PCR (Figure 3). IL28B expression was approximately 10-fold less than the expression of ISGs. Although IL28B expression tended to be higher in some patients with the major genotype, there was no significant difference in IL28B expression in the liver between the major and minor genotypes (Figure 3*A*). Nevertheless, the expression of ISGs was clearly high in patients with the minor genotype ($P < .0001$) (Figure 3*B*). IL28 activates signal tranducers and activators of transcription 1 (STAT1) through downstream signaling from a heterodimeric class II cytokine receptor that consists of IL-10 receptor β (IL-10Rβ) and IL-28 receptor α (IL-28Rα).[18,19] Therefore, we examined the correlation between the expression of IL28B and ISGs. IL28B expression correlated with the expression of ISGs ($r = 0.44$, $P < .001$); however, the correlation was different according to the SNP genotype. We observed a steep-slope correlation for the minor genotype and a slow-slope correlation for the major genotype (Figure 3*C* and *D*). Interestingly, 4 minor homozygotic (GG) patients showed a steeper correlation than the heterozygotes (TG) (Figure 3*D*). Thus, the IL28B polymorphism might differentially regulate the expression of ISGs in the liver, leading to the different treatment outcomes.

We performed univariate and multivariate analyses to identify the clinical factors associated with the major and minor genotypes (Table 4). Univariate analysis showed that higher hepatic ISGs and lower body mass index were significantly associated with the minor genotype; however, multivariate analysis showed that only hepatic ISGs (≥3.5) were associated with the minor genotype ($P < .001$; OR, 18.1; 95% confidence interval: 3.95–113). We further compared the predictive capacity of multivariate models using the expression of hepatic ISGs (<3.5 vs ≥3.5) or the IL28B genotype (major vs minor) (Supplementary Table 6). The predictive performance and fitness of the multivariate model using the IL28B genotype was superior to that using the expression of hepatic ISGs. However, when these factors were included in the same model, the expression of hepatic ISGs was still useful for the predictive model independent of the IL28B genotype (Supplementary Table 6).

Table 3. Clinical Characteristics of 32 Patients Genotyped by GWAS and 5 SNPs in Strong Linkage Disequilibrium With IL28B,[11] Including rs12979860

Patient No.	Response	Age, y	Sex	F stage	ISGs	IL28B	RefSNP (chr pos) Minor allele	rs8105790 (44424341) C	rs11881222 (44426763) G	rs12979860 (44430627) T	rs8099917 (44435005) G	rs7248668 (44435661) A
1	SVR	42	M	1	4.20	83.8		TT	AA	CC	TT	GG
2	SVR	59	M	1	2.62	45.5		TT	AA	CC	TT	GG
3	SVR	41	F	1	1.54	1.3		TT	AA	CC	TT	GG
4	TR	57	M	1	3.18	21.7		TT	AA	CC	TT	GG
5	TR	68	F	1	1.43	20.3		TT	AA	CC	TT	GG
6	SVR	44	M	1	0.97	4.6		TT	AA	CC	TT	GG
7	SVR	61	M	2	2.15	6.1		TT	AA	CC	TT	GG
8	SVR	50	M	2	3.25	66.4		TT	AA	CC	TT	GG
9	SVR	49	M	2	1.25	ND		TT	AA	CC	TT	GG
10	TR	59	F	2	1.29	17.4		TT	AA	CC	TT	GG
11	SVR	48	F	2	1.00	90.2		TT	AA	CC	TT	GG
12	TR	65	F	2	2.86	36.4		TT	AA	CC	TT	GG
13	NR	34	M	3	0.82	17.8		TT	AA	CC	TT	GG
14	SVR	55	M	3	0.83	13.8		TT	AA	CC	TT	GG
15	TR	68	M	3	0.75	20.6		TT	AA	CC	TT	GG
16	SVR	64	M	3	0.94	15.7		TT	AA	CC	TT	GG
17	SVR	67	F	3	1.50	25.7		TT	AA	CC	TT	GG
18	SVR	48	M	4	1.69	7.9		TT	AA	CC	TT	GG
19	NR	66	F	1	4.57	16.5		TC	AG	CT	TG	GA
20	SVR	52	F	1	5.23	29.3		TC	AG	CT	TG	GA
21	NR	55	F	1	8.25	57.2		TC	AG	CT	TG	GA
22	SVR	49	F	1	5.36	ND		TC	AG	CT	TG	GA
23	TR	44	M	1	2.08	7.0		TC	AG	CT	TG	GA
24	NR	63	M	1	2.77	10.5		TC	AG	CT	TG	GA
25	NR	61	F	2	3.98	39.1		TC	AG	CT	TG	GA
26	NR	42	M	2	4.89	5.9		TC	AG	CT	TG	GA
27	SVR	49	M	3	3.31	6.9		TC	AG	CT	TG	GA
28	TR	71	F	3	5.53	27.3		TC	AG	CT	TG	GA
29	TR	63	M	3	3.40	33.5		TC	AG	CT	TG	GA
30	NR	70	F	3	4.78	8.1		TC	AG	CT	TG	GA
31	TR	62	F	3	3.53	14.0		TC	AG	CT	TG	GA
32	NR	56	M	4	7.37	30.8		CC	GG	TT	GG	AA

NOTE. The Pearson correlation of the r^2 estimates for adjacent pairs; rs8099917 vs rs8105790, rs8099917 vs rs11881222, rs8099917 vs rs12979860, and rs8099917 vs rs7248668 = 0.99, 0.99, 0.98, and 0.97, respectively.

IL28B, interleukin 28B; GWAS, genome-wide association studies; ISGs, interferon stimulated genes; SNP, single nucleotide polymorphism; SVR, sustained viral response; TR, transient response; NR, no response; M, male; F, female.

To examine further the different hepatic gene expression of patients with the major or minor genotypes, pathway analysis of differentially expressed genes between the 2 groups was performed. By comparing the expression of hepatic genes between patients with the major and minor genotypes, 1359 differentially expressed genes were identified ($P < .01$; 711 genes were up-regulated with the minor genotype, and 648 genes were up-regulated with the major genotype). Pathway analysis of these genes demonstrated that signaling pathways related to interferon action, apoptosis, and Wnt signaling were up-regulated in the liver of patients with the minor genotype, whereas B-cell-, dendritic cell-, and natural killer cell-related genes were up-regulated in the liver of patients with the major genotype (Supplementary Figure 3). These results suggest that IL28B may be involved in innate and adaptive immune responses and that different antiviral signaling pathways might be involved in the liver of patients with different SNPs.

Discussion

Multiple viral and host factors may be related to the treatment response to Peg-IFN and RBV combination therapy. For the viral factors, a higher number of aa substitutions in the ISDR of nonstructural 5A region was strongly associated with a favorable response to IFN-α monotherapy in patients with genotype-1 HCV.[4]

Besides viral factors, host factors such as age, gender, fibrotic stage of the liver, and the presence of steatosis and insulin resistance were associated with the treatment outcome.[20] Analysis of hepatic gene expression demonstrated that the up-regulation of ISGs in the liver before treatment may be related to a poor treatment response.[6-9] To reveal the underlying mechanism of treatment resistance, 2 reports compared gene expression profiling in the liver before and during therapy and showed that patients with up-regulated ISGs in the liver prior to treatment failed to further induce ISGs following the ad-

Table 4. Comparison of Clinical Factors Between Patients With Major (TT) and Minor (TG+GG) Alleles

Clinical category	TT		TG+GG		Univariate *P* value	Multivariate odds (95% CI)	Multivariate *P* value
No. of patients	n = 60		n = 31			—	
Treatment response							
SVR+TR vs NR	51 vs 9		11 vs 20		<.001	—	
Age and gender							
Age, *y*	56	(30–69)	56	(30–71)	.843	—	
Sex (M vs F)	39 vs 21		19 vs 12		.518	—	
Liver factors							
F stage (F1-2 vs F3-4)	36 vs 24		23 vs 17		.905	—	
A grade (A0-1 vs A2-3)	27 vs 33		20 vs 11		.075	—	
ISGs (Mx, IFI44, IFIT1)							
(<3.5 vs ≥3.5)	46 vs 14		5 vs 26		<.001	18.1 (3.95–113)	<.001
Laboratory parameters							
HCV-RNA (*KIU/mL*)	2055	(160–5000)	1970	(126–5000)	.602	—	
BMI (kg/m^2)	24.5	(16.3–40.5)	22.9	(19.1–26.6)	.006	—	.077
AST (*IU/L*)	59	(22–258)	54	(21–283)	.227	—	
ALT (*IU/L*)	75	(24–376)	60	(18–236)	.077	—	
γ-GTP (*IU/L*)	61	(4–392)	53	(20–229)	.517	—	.167
WBC ($/mm^3$)	4450	(2100–11,100)	4600	(2500–8200)	.947	—	
Hb (*g/dL*)	14.2	(11.4–16.7)	14.5	(11.2–17.2)	.606	—	
PLT ($\times 10^4/mm^3$)	15.4	(7–39.4)	16.2	(9.2–27.7)	.832	—	
TG (*mg/dL*)	98	(58–248)	131	(30–303)	.053	—	.055
T-Chol (*mg/dL*)	172	(115–222)	168	(129–237)	.910	—	
LDL-Chol (*mg/dL*)	84	(42–123)	69	(51–107)	.052	—	.055
HDL-Chol (*mg/dL*)	44	(18–72)	45	(29–77)	.218	—	
FBS (*mg/dL*)	95	(59–291)	96	(66–206)	.849	—	
Insulin (*μU/mL*)	7.5	(0.7–23.2)	9.2	(2–23.2)	.195	—	
HOMA-IR	1.3	(0.3–11.7)	1.2	(0.4–9.6)	.339	—	
Viral factors							
ISDR mutations (≤1 vs ≥2)	38 vs 22		23 vs 7		.194	—	.083
Treatment factors							
Total dose administrated							
Peg-IFN (μg)	3960	(1500–7200)	3840	(1920–5760)	.377	—	
RBV (g)	203	(26–336)	201	(106–268)	.777	—	
Achieved administration rate							
Peg-IFN (%)							
≥80%	41		17		.207	—	
<80%	19		14				
RBV (%)							
≥80%	34		19		.671	—	
<80%	26		12				
Achievement of EVR	40/60 (62%)		9/31 (29%)		<.001	—	

BMI, body mass index; AST, aspartate aminotransferase; ALT, alanine aminotransferase; IFI44, interferon-induced protein 44; IFIT1, interferon-induced protein with tetratricopeptide repeats 1; EVR, early virologic response; γ-GTP, γ-glutamyl transpeptidase; ISDR, interferon sensitivity determining region; Mx1, myxovirus (influenza virus) resistance 1 interferon-inducible protein p78 (mouse); WBC, leukocytes; HOMA-IR, homeostasis model assessment of insulin resistance; Hb, hemoglobin; RBV, ribavirin; PLT, platelets; TG, triglycerides; TR, transient response; T-chol, total cholesterol; LDL-chol, low-density lipoprotein cholesterol; HDL-chol, high-density lipoprotein cholesterol; FBS, fasting blood sugar; CI, confidence interval.

ministration of IFN and could not eliminate HCV.[6,7] We performed a similar analysis and observed that these findings were more evident in liver lobular cells than in infiltrating lymphocytes in the portal area (submitted for publication). Thus, both viral and host factors might be closely related to the treatment response to Peg-IFN and RBV combination therapy. However, the clinical relevance and relationships of these factors have not been fully evaluated. In this study, we validated the clinical significance of the expression of hepatic ISGs on treatment outcome using a relatively large cohort of patients and compared its significance with other viral and host factors. To compare the patients with SVR, TR, and NR, we assessed the overall plausibility of each group using Fisher C statistic,[16] and patients with SVR and TR were grouped together for further analysis.

We examined hepatic gene expression in 91 of 168 patients using the Affymetrix genechip. Expression profiling using 37 representative ISGs (see Materials and Methods), which were selected from gene expression profiling comparing pretreatment and under treatment liver, differentiated 2 groups of

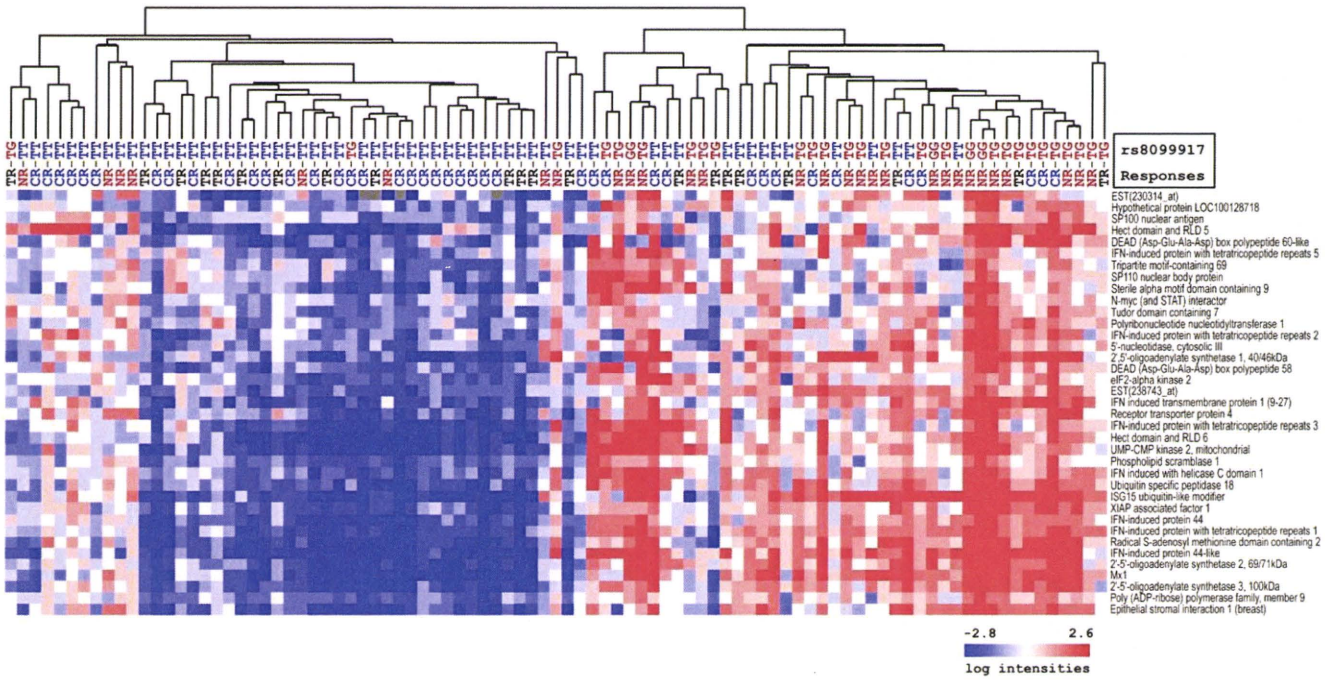

Figure 2. Hierarchical clustering analysis of 91 patients with the defined genotype of IL28B. Responses to therapy (SVR, TR, and NR) and IL28B genotype (TT, TG, or GG) are shown. The structure of the dendrogram and heat map is the same as in Figure 1.

patients: the Up-ISG and Down-ISG groups (Figure 1). The proportion of patients with NR to treatment was significantly higher in the Up-ISGs group.

Multivariate analysis showed that hepatic ISGs (<3.5), fibrosis stage (F1-F2), and ISDR mutations (≥2) significantly contributed to the outcome for the SVR+TR group (Table 1). Discriminate analysis using variables selected by multivariable analysis predicted the SVR+TR patients with 82% accuracy and NR patients with 79% accuracy. However, the accuracy decreased to 67% for SVR+TR patients and 53% for NR patients when the expression of hepatic ISGs was removed from the variables (data not shown). Interestingly, the expression of hepatic ISGs was strongly correlated with γ-GTP and weakly correlated with insulin resistance. A recent study describing the association between insulin resistance and poor treatment outcome might be partially explained by this observation.[20]

In this study, we utilized 3 ISGs (Mx1, IFI44, and IFIT1) out of 15 validated by RTD-PCR. The expression values of these ISGs were higher than those of other ISGs (Supplementary Figure 1*A*). We averaged these ISGs and set the cut-off value as 3.5 from the ROC curve (Supplementary Figure 1*B*). The sensitivity, specificity, and positive and negative predictive values on the likelihood of achieving SVR+TR using this cut-off value were 82% (103/125), 72% (31/43), 90% (103/115), and 58% (31/53), respectively. The results were compared with those observed for the 15 ISGs (Supplementary Table 5). These results showed that the 3.5 cut-off value for Mx1, IFI44, and IFIT1 would be valuable for clinical use.

Despite the importance of the expression of hepatic ISGs, viral factors may also allow us to predict the outcome of treatment. Multivariate analysis showed that ISDR mutations (≥2) independently contributed to the treatment outcome, although univariate analysis did not show significance (P = .07); therefore, ISDR might be uniquely and differentially involved in treatment resistance.

What causes the differences in the expression of hepatic ISGs? In parallel to the gene expression analysis, a GWAS was applied to identify genomic loci associated with treatment response, and a polymorphism in IL28B was found to predict hepatitis C treatment-induced viral clearance.[10–12] To examine the relationship between the genetic variation of IL28B and hepatic gene expression, we determined the IL28B polymorphism in 91 patents (Table 3). The patients with the minor genotype (TG or GG) had an increased expression of hepatic ISGs compared with the patients with major genotype (TT) (Figures 2 and 3). In European-Americans, the proportion of major homozygotes is 39% (CC at rs1297986), 49% for heterozygotes (TC), and 12% for minor homozygotes (TT).[11] Although the proportion of minor homozygotes was much less in this study (GG, 4%), as reported in a previous study in Japan,[12] more patients are required for proper evaluation. It is interesting that the expression of hepatic ISGs in minor homozygotes (GG) was higher than in heterozygotes (TG) in this study.

The results clearly showed that the differences in the expression of hepatic ISGs before treatment are associated with the IL28B polymorphism and results in different treatment outcomes. Although we could not detect significant differences in the expression levels of IL28B depending on the different SNP, some patients with the major genotype showed a higher expression of IL28B. Because IL28B expression was approximately 10-fold less than the expression of ISGs, the lower

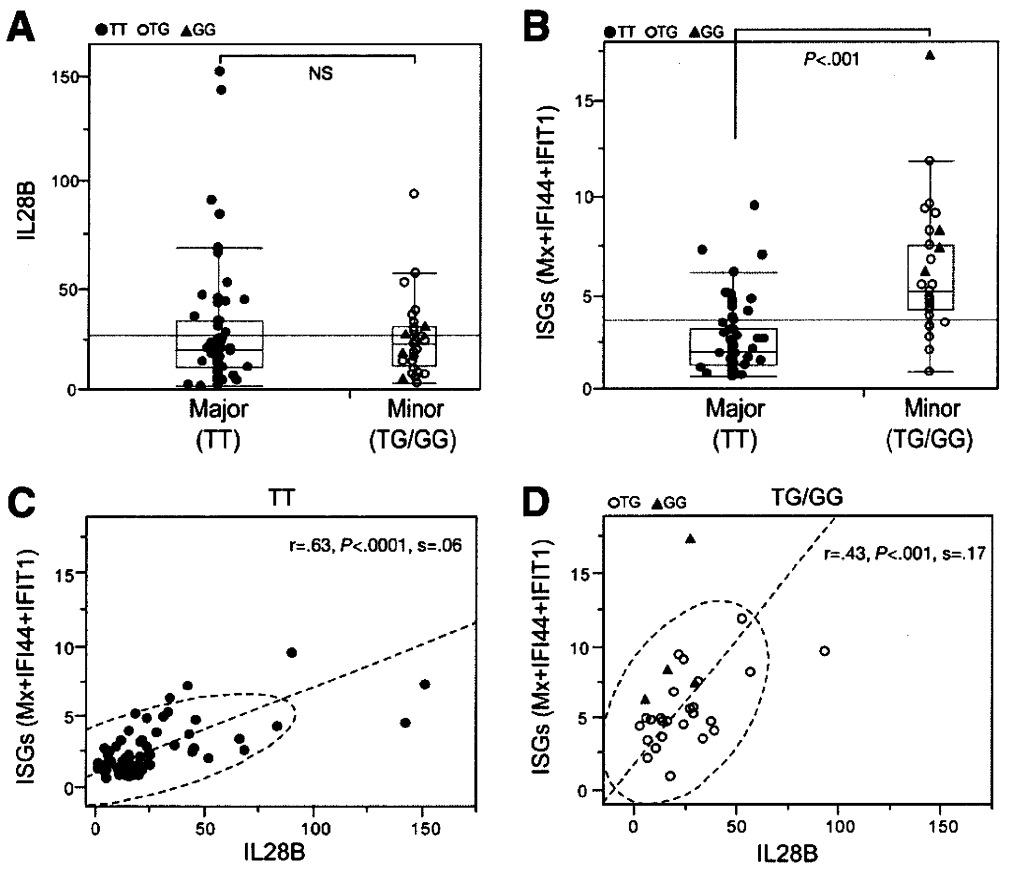

Figure 3. (*A*) IL28 expression in the liver of 91 patients with the major (TT) or minor (TG or GG) genotype (rs8099917). (*B*) Expression of ISGs in the liver of patients with the major (TT) or minor (TG or GG) genotype (rs8099917). (*C*) Relationship between IL28 and ISGs in the liver of patients with the major (TT) genotype (rs8099917). (*D*) Relationship between IL28 and ISGs in the liver of patients with the minor (TG or GG) genotype (rs8099917).

expression of IL28B may be a reason for the decreased ability to distinguish differences in its expression. Another possibility may be the specificity of the IL28B primers used in this study; because IL28B shares a 98.2% nucleotide sequence homology with IL28A, IL28B specific primers are not available.[21] When the expression of IL28B and hepatic ISGs were compared, a significant correlation was observed, and, interestingly, IL28B and ISGs derived from different SNPs were correlated in a different way (Figure 3*C* and *D*). It appeared that hepatic ISGs were more induced by the reduced amounts of IL28B in patients with the minor genotype. The mechanism behind these findings has yet to be determined; however, IL28B interacts with a heterodimeric class II cytokine receptor that consists of IL-10 receptor β (IL-10Rβ) and IL-28 receptor α (IL-28Rα).[18,19] It is possible that IL28B could mediate antiviral signaling through IL-10 signaling as well as STAT1 activation. The Th 2 dominant signaling of IL28B may modulate signaling pathways in livers with CH-C and contributes to the different expression of ISGs. Another possibility may be that the cell origin of hepatic ISGs is different. A recent study revealed cell-type specific ISG expression in macrophages and hepatocytes, which could be related to the IFN response.[22] As more of the B-cell-, dendritic cell-, and natural killer cell-related genes were up-regulated in the liver of patients with the major genotype, ISGs could be expressed by these cells, whereas they are expressed by hepatocytes in the liver of patients with the minor genotype. It is known that the induction of ISGs in lymphocytes is lower than that in hepatocytes. The precise mechanism should be investigated further as a different regulatory mechanism for the expression of ISGs may be present.

In conclusion, we presented the clinical relevance of the expression of hepatic ISGs for the treatment outcome of Peg-IFN and RBV combination therapy. The different expressions of hepatic ISGs before treatment might be due to polymorphisms in IL28B. Further studies are required to clarify the detailed pathways of IL28B and hepatic gene expression through molecular biologic and immunologic aspects.

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.2010.04.049.

Appendix 1. The Hokuriku Liver Study Group (HLSG) is Composed of the Following Members:

Drs Takashi Kagaya, Kuniaki Arai, Kaheita Kakinoki, Kazunori Kawaguchi, Hajime Takatori, Hajime Sunakosaka (Department of Gastroenterology, Kanazawa University Graduate School of Medicine, Kanazawa); Drs Touru

Nakahama, Shinji Kamiyamamoto (Kurobe City Hospital, Kurobe, Toyama); Dr Yasuhiro Takemori (Toyama Rosai Hospital, Uozu, Toyama); Dr Hikaru Oguri (Koseiren Namerikawa Hospital, Namerikawa, Toyama); Drs Yatsugi Noda, Hidero Ogino (Toyama Prefectural Central Hospital, Toyama, Toyama); Drs Yoshinobu Hinoue, Keiji Minouchi (Toyama City Hospital, Toyama, Toyama); Dr Nobuyuki Hirai (Koseiren Takaoka Hospital, Takaoka, Toyama); Drs Tatsuho Sugimoto, Koji Adachi (Tonami General Hospital, Tonami, Toyama); Dr Yuichi Nakamura (Noto General Hospital, Nanao, Ishikawa); Drs Masashi Unoura, Ryuhei Nishino (Public Hakui Hospital, Hakui, Ishikawa); Drs Hideo Morimoto, Hajime Ohta (National Hospital Organization Kanazawa Medical Center, Kanazawa, Ishikawa); Dr Hirokazu Tsuji (Kanazawa Municipal Hospital, Kanazawa, Ishikawa); Drs Akira Iwata, Shuichi Terasaki (Kanazawa Red Cross Hospital, Kanazawa, Ishikawa); Drs Tokio Wakabayashi, Yukihiro Shirota (Saiseikai Kanazawa Hospital, Kanazawa, Ishikawa); Drs Takeshi Urabe, Hiroshi Kawai (Public Central Hospital of Matto Ishikawa, Hakusan, Ishikawa); Dr Yasutsugu Mizuno (Nomi Municipal Hospital, Nomi, Ishikawa); Dr Shoni Kameda (Komatsu Municipal Hospital, Komatsu, Ishikawa); Drs Hirotoshi Miyamori, Uichiro Fuchizaki (Keiju Medical Center, Nanao, Ishikawa); Dr Haruhiko Shyugo (Kanazawa Arimatsu Hospital, Kanazawa, Ishikawa); Dr Hideki Osaka (Yawata Medical Center, Komatsu, Ishikawa); Dr Eiki Matsushita (Kahoku Central Hospital, Tsubata, Ishikawa); Dr Yasuhiro Katou (Katou Hospital, Komatsu, Ishikawa); Drs Nobuyoshi Tanaka, Kazuo Notsumata (Fukuiken Saiseikai Hospital, Fukuil, Fukui); Dr Mikio Kumagai (Kumagai Clinic, Tsuruga, Fukui); Dr Manabu Yoneshima (Municipal Tsuruga Hospital, Tsuruga, Fukui).

References

1. Kiyosawa K, Sodeyama T, Tanaka E, et al. Interrelationship of blood transfusion, non-A, non-B hepatitis and hepatocellular carcinoma: analysis by detection of antibody to hepatitis C virus. Hepatology 1990;12:671–675.
2. Fried MW, Shiffman ML, Reddy KR, et al. Peginterferon alfa-2a plus ribavirin for chronic hepatitis C virus infection. N Engl J Med 2002;347:975–982.
3. Poynard T, Ratziu V, McHutchison J, et al. Effect of treatment with peginterferon or interferon alfa-2b and ribavirin on steatosis in patients infected with hepatitis C. Hepatology 2003;38:75–85.
4. Enomoto N, Sakuma I, Asahina Y, et al. Mutations in the nonstructural protein 5A gene and response to interferon in patients with chronic hepatitis C virus 1b infection. N Engl J Med 1996; 334:77–81.
5. Okanoue T, Itoh Y, Hashimoto H, et al. Predictive values of amino acid sequences of the core and NS5A regions in antiviral therapy for hepatitis C: a Japanese multi-center study. J Gastroenterol 2009;44:952–963.
6. Feld JJ, Nanda S, Huang Y, et al. Hepatic gene expression during treatment with peginterferon and ribavirin: identifying molecular pathways for treatment response. Hepatology 2007;46:1548–1563.
7. Sarasin-Filipowicz M, Oakeley EJ, Duong FH, et al. Interferon signaling and treatment outcome in chronic hepatitis C. Proc Natl Acad Sci U S A 2008;105:7034–7039.
8. Asselah T, Bieche I, Narguet S, et al. Liver gene expression signature to predict response to pegylated interferon plus ribavirin combination therapy in patients with chronic hepatitis C. Gut 2008;57:516–524.
9. Chen L, Borozan I, Feld J, et al. Hepatic gene expression discriminates responders and nonresponders in treatment of chronic hepatitis C viral infection. Gastroenterology 2005;128:1437–1444.
10. Thomas DL, Thio CL, Martin MP, et al. Genetic variation in IL28B and spontaneous clearance of hepatitis C virus. Nature 2009; 461:798–801.
11. Ge D, Fellay J, Thompson AJ, et al. Genetic variation in IL28B predicts hepatitis C treatment-induced viral clearance. Nature 2009;461:399–401.
12. Tanaka Y, Nishida N, Sugiyama M, et al. Genome-wide association of IL28B with response to pegylated interferon-α and ribavirin therapy for chronic hepatitis C. Nat Genet 2009;41:1105–1109.
13. Desmet VJ, Gerber M, Hoofnagle JH, et al. Classification of chronic hepatitis: diagnosis, grading and staging. Hepatology 1994;19:1513–1520.
14. Honda M, Yamashita T, Ueda T, et al. Different signaling pathways in the livers of patients with chronic hepatitis B or chronic hepatitis C. Hepatology 2006;44:1122–1138.
15. Honda M, Kaneko S, Kawai H, et al. Differential gene expression between chronic hepatitis B and C hepatic lesion. Gastroenterology 2001;120:955–966.
16. Shipley B. A new inferential test for path models based on directed acyclic graphs. Structural Equation Modeling 2000;7:206–218.
17. Favre M, Martin J, Festa-Bianchet M. Determinants and life-history consequences of social dominance in bighorn ewes. Anim Behav 2008;76:1373–1380.
18. Sheppard P, Kindsvogel W, Xu W, et al. IL-28, IL-29, and their class II cytokine receptor IL-28R. Nat Immunol 2003;4:63–68.
19. Kotenko SV, Gallagher G, Baurin VV, et al. IFN-λ mediate antiviral protection through a distinct class II cytokine receptor complex. Nat Immunol 2003;4:69–77.
20. Camma C, Bruno S, Di Marco V, et al. Insulin resistance is associated with steatosis in nondiabetic patients with genotype 1 chronic hepatitis C. Hepatology 2006;43:64–71.
21. Mihm S, Frese M, Meier V, et al. Interferon type I gene expression in chronic hepatitis C. Lab Invest 2004;84:1148–1159.
22. Chen L, Borozan I, Sun J, et al. Cell-type specific gene expression signature in liver underlies response to interferon therapy in chronic hepatitis C infection. Gastroenterology 2010;138:942–948.

Received October 9, 2009. Accepted April 14, 2010.

Reprint requests

Address requests for reprints to: Shuchi kaneko, MD, PhD, Department of Gastroenterology, Graduate School of Medicine, Kanazawa University, Takara-Machi 13-1, Kanazawa 920-8641, Japan. e-mail: skaneko@m-kanazawa.jp; fax: (81) 76-234-4250.

Acknowledgments

The authors thank Nami Nishiyama and Yuki Hatayama for excellent technical assistance.

Participating investigators are listed in Appendix 1.

Conflicts of Interest

The authors disclose no conflicts.

Funding

This work was supported in part by a grant-in-aid from the Ministry of Health, Labour and Welfare of Japan.

Nakamura et al. BMC Bioinformatics 2010, 11:398
http://www.biomedcentral.com/1471-2105/11/398

BMC Bioinformatics

DATABASE Open Access

CIG-DB: the database for human or mouse immunoglobulin and T cell receptor genes available for cancer studies

Yoji Nakamura[1,5], Tomoyoshi Komiyama[2], Motoki Furue[3], Takashi Gojobori[4], Yasuto Akiyama[1*]

Abstract

Background: Immunoglobulin (IG or antibody) and the T-cell receptor (TR) are pivotal proteins in the immune system of higher organisms. In cancer immunotherapy, the immune responses mediated by tumor-epitope-binding IG or TR play important roles in anticancer effects. Although there are public databases specific for immunological genes, their contents have not been associated with clinical studies. Therefore, we developed an integrated database of IG/TR data reported in cancer studies (the Cancer-related Immunological Gene Database [CIG-DB]).

Description: This database is designed as a platform to explore public human and murine IG/TR genes sequenced in cancer studies. A total of 38,308 annotation entries for IG/TR proteins were collected from GenBank/DDBJ/EMBL and the Protein Data Bank, and 2,740 non-redundant corresponding MEDLINE references were appended. Next, we filtered the MEDLINE texts by MeSH terms, titles, and abstracts containing keywords related to cancer. After we performed a manual check, we classified the protein entries into two groups: 611 on cancer therapy (Group I) and 1,470 on hematological tumors (Group II). Thus, a total of 2,081 cancer-related IG and TR entries were tabularized. To effectively classify future entries, we developed a computational method based on text mining and canonical discriminant analysis by parsing MeSH/title/abstract words. We performed a leave-one-out cross validation for the method, which showed high accuracy rates: 94.6% for IG references and 94.7% for TR references. We also collected 920 epitope sequences bound with IG/TR. The CIG-DB is equipped with search engines for amino acid sequences and MEDLINE references, sequence analysis tools, and a 3D viewer. This database is accessible without charge or registration at http://www.scchr-cigdb.jp/, and the search results are freely downloadable.

Conclusions: The CIG-DB serves as a bridge between immunological gene data and cancer studies, presenting annotation on IG, TR, and their epitopes. This database contains IG and TR data classified into two cancer-related groups and is able to automatically classify accumulating entries into these groups. The entries in Group I are particularly crucial for cancer immunotherapy, providing supportive information for genetic engineering of novel antibody medicines, tumor-specific TR, and peptide vaccines.

Background

The immune system is inherent in vertebrates and provides protection against toxic substances or infectious diseases. Two antigen receptor proteins, immunoglobulin (IG) expressed on B lymphocytes or secreted by plasma cells, and the T-cell receptor (TR), expressed on T lymphocytes, are key molecules for humoral immunity and cell-mediated immunity, respectively [1]. Each of these proteins consists of two chain types, called light and heavy chains for IG (there are two identical light chains and two identical heavy chains in an IG), and alpha and beta chains, or gamma and delta chains for TR. Each chain contains, at its N-terminal end, a variable (V) domain which participates in antigen recognition. The V domain is encoded by two or three genes, a V gene, a diversity (D) gene (for heavy, beta and delta chains) and a joining (J) gene, which rearrange through somatic recombination [2]. In the V domain, three complementarity determining regions (CDRs), which are especially sequence-diversified, contact antigenic epitopes. In particular, the third CDR (CDR3) is the most

* Correspondence: y.akiyama@scchr.jp
[1]Immunotherapy Division, Shizuoka Cancer Center Research Institute, 1007 Shimonagakubo, Nagaizumi-cho, Sunto-gun, Shizuoka, 411-8777, Japan

diversified among the CDRs at the junction of V(D)J recombination and is considered crucial for the recognition of epitopes [3-5].

Cancer cells proliferate abnormally compared to normal cells, often expressing proteins (tumor-associated antigens) that cannot be seen in normal developmental stages [6]. In cancer studies, monitoring the immune status of patients is thus very important for diagnosis, as expression of an autoantibody [7] and the activation of cytotoxic T lymphocytes (CTLs) [8] specific to tumor-associated antigens are observed. In hematological tumors, such as leukemia or lymphoma, IG and TR themselves are the subject of investigation, because the encoding genes are often mutated by translocation in tumor B or T cells [9]. Moreover, in recent years, these antigen receptor proteins have attracted attention in the field of cancer immunotherapy to elevate the patient's immune response against tumor cells with few side-effects [10,11]. In cellular immunotherapy, T cells recognizing tumor-associated antigens can be administrated back to patients after *ex vivo* culture and processing for immune response enhancement.

During the last decade, monoclonal antibodies have been sought and engineered as candidates for molecular target drugs [12]. These molecules can recognize the cancer cells expressing tumor-associated antigens with high affinity and selectivity, triggering anticancer effects [12,13], such as complement dependent cytotoxicity, antibody-dependent cellular cytotoxicity, inhibition of angiogenesis, and induction of apoptosis. In general, the source of antibody medicines is the human or mouse: (i) fully murine, (ii) chimeric with V domains from the mouse and constant regions from the human, and (iii) humanized or human antibodies have been developed [12]. For instance, trastuzumab (trade name Herceptin), a humanized antibody that targets the human epidermal growth factor receptor type 2 protein, has shown success in the treatment of breast cancer [14]. Around 10 antibody medicines in cancer therapy are now approved and another 30 are being assessed in clinical trials in the USA [12].

Cancer vaccines are another type of immunotherapy, in which partial epitope peptides of tumor-associated antigens are administered to patients to potentiate CTL activity [15]. TR on CD8+ T cells recognize the peptide vaccines bound with human leukocyte antigens (HLA), which are displayed by antigen-presenting cells, and enhance cytotoxic activity against cancer cells carrying the peptides [16]. Each peptide vaccine is, in general, 9-10 amino acids long and selected according to the patient's HLA allele type.

These immunotherapeutic studies have emphasized the importance of genetic engineering of IG and TR proteins and peptide vaccines at the sequence level [17-21]. Currently, there are several immunological gene databases available online, such as IMGT®, the international ImMunoGeneTics information system® [22] and the Immune Epitope Database and Analysis Resource (IEDB) [23]. Although the contents of these databases are well-annotated and specific to genetic information on IG, TR, and their epitopes, there are still gaps between this information and the clinical application in cancer research. In particular, the information supplied is too broad for clinicians and pharmacologists (for example, such databases store the data from a large variety of organisms and the majority are unrelated to cancer) to easily obtain the information required for patient-specific treatment. To address these issues, we have developed a freely accessible database, the Cancer-related Immunological Gene Database (CIG-DB). The database integrates the information on IG, TR, and epitopes reported in cancer studies, and presents sequence analysis tools, and structural data.

Construction and content

Data integration

The CIG-DB is a semi-automated database consisting of four tables, two of which are for IG and TR proteins, respectively. All the included proteins are only derived from the human or mouse. The other two tables are for epitopes of IG or TR, where we collected the amino acid sequences from a variety of organisms regardless of whether they were cancer-related or not. Since the available amount of cancer-related epitope data is still small, a large number of sequences are considered useful for further comparative analysis.

First, to thoroughly check all public IG/TR proteins of human and mouse origins, we downloaded their annotation data from GenBank/DDBJ/EMBL by keyword search. Then, the Geninfo Identifier (GI) and accession numbers, full amino acid sequences, and references were extracted and tabularized as a local proto-database. CDR3 sequences were also extracted from the full sequences using a pattern-matching program to find the N- and C-terminal borders. The program detects conserved motifs flanking the CDR3 region using regular expressions, such as the second cysteine and its adjacent residues as the N-flanking sequence and a WGXG motif as the C-flanking sequence. The V and J gene repertoires were assigned according to IMGT® nomenclature [24,25] and BLAST [26] matching. Furthermore, we obtained structural data of the IG/TR proteins from the Protein Data Bank (PDB) [27], if any, and the information was merged with the GenBank annotation. We thus obtained a total of 32,240 entries for IG and 6,068 for TR, as of October 1, 2009 (Table 1). Next, for each of the entries, we obtained the corresponding 2,740 MEDLINE reference data from PubMed at the National

Table 1 CIG-DB statistics as of 1 October 2009

Data content	IG	TR	Total
Screened from NCBI and PDB	32240	6068	38308
Cancer-related	1605	476	2081
Human	879	318	1197
Mouse	726	158	884
Group I[a]	397	214	611
Group II[b]	1208	262	1470
Chains			
Light	791		
Heavy	814		
Alpha		185	
Beta		288	
Gamma[c]		3	
Epitope sqeuences	772	148	920

[a]This group is involved in cancer therapy such as antibody medicines.
[b]This group is involved in expression in hematological tumors.
[c]A minor TCR allele.

Center for Biotechnology Information (Table 2). The number of references is approximately one order of magnitude smaller than that of the receptor sequences, because multiple IG or TR sequences are often reported in a single reference. We then selected 578 MEDLINE reference texts in which cancer-related keywords ("melanoma," "carcinoma," "sarcoma," etc.) are contained in at least one of either the MeSH terms, title, or abstract sentences. When we annotated the IG/TR entries according to the selected references, we noticed that the IG/TR entries screened were used for the following cases in cancer studies: (i) studied in the context of cancer therapy (specific TR and antibodies monitored after administration of peptide vaccines in immunotherapy, monoclonal antibodies specific to cancer markers, etc.) and (ii) isolated from hematological tumors such as leukemia, lymphoma, and myeloma (survey of mutation

Table 2 Classification of cancer-related references in CIG-DB

Data content	IG	TR	Total
Collected from PubMed	2054	686	2740
Screened by keywords	446	132	578
Manually classified[a]			
Group I	120	34	
Group II	139	37	
(Group III)	187	61	

[a]Groups I and II are the same as in Table 1. Group III is a group of wrongly screened references.

in IG or TR gene locus, etc.). Therefore, we manually classified the IG/TR entries into these two groups, named Groups I (cancer therapy) and II (hematological tumors), and the corresponding entries were tabularized in the same format as the above-mentioned proto-database. Finally, we obtained a total of 2,081 cancer-related receptor entries, 1,605 for IG and 476 for TR (Table 1).

Regarding epitopes, the interaction between IG/TR and an epitope is the specific focus of this database. The source of epitope sequences was public databases: the IEDB, Bcipep [28], and the HIV sequence database [29], and where possible the PDB, if epitopes were crystallized as bound with IG or TR. We then extracted the protein or peptide epitopes and selected those whose antigen receptor sequences are known. The matching criteria were as follows: (i) the complex structure of the antigen receptor and epitope are already in the PDB, (ii) for the epitopes from the IEDB, the GI numbers of the receptors are found in the IG/TR tables of the CIG-DB, and (iii) manual check of references. As a whole, we obtained a total of 920 epitope sequences, 772 for IG, and 148 for TR (Table 1). Antigen-presenting cells display as T cell epitopes around 9-mer peptides, and TR epitopes are therefore always of linear sequences. In the case of IG epitopes, conformational ones neighbored through antigen folding are possible. To show the key residues involved in the interaction with the antigen receptor, residues were highlighted based on the distances of the residues to the receptor's CDR3 (< 4 angstroms).

Reference clustering and classification

In the current version of the CIG-DB, as mentioned above, we manually checked the references to classify the IG and TR entries into two cancer-related groups. We further prepared an automated classification method based on text mining, training, and clustering algorithms. For training, we appended one additional group (Group III) of unrelated references (Table 2), which was wrongly selected by cancer-related keywords and dropped by manual classification. First, we parsed each of the MEDLINE text format data of the three groups using the *tm* library of the R program and computed a term frequency vector for the words occurring in the MeSH terms, title, and abstract, after processing and stemming. Figure 1 shows the result of principal component analysis (PCA), which plots the first and second principal components (PC1 and PC2) using all the term frequency vectors, suggesting that the three groups are distributed separately from each other for both IG and TR references. Next, we averaged and merged all the frequency vectors of the references in each group and calculated a canonical discriminant function, using the Mahalanobis distance between any

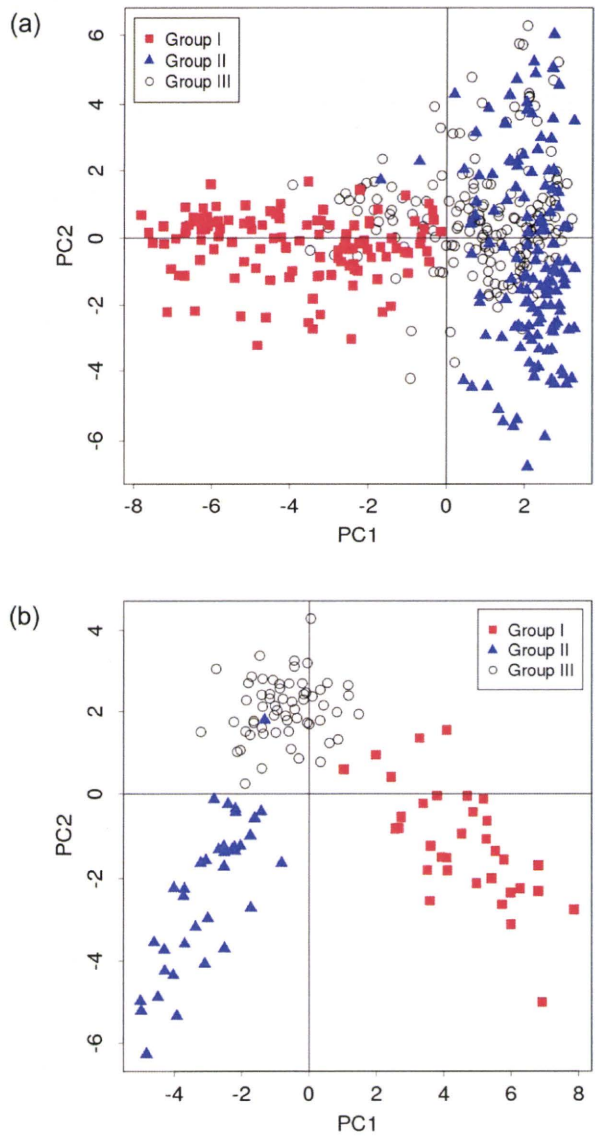

Figure 1 The result of principal component analysis on cancer-related immunoglobulin (IG) and T-cell receptor (TR) references. PC1 and PC2 indicate scores of the first and second principal components, respectively. (a) IG references and (b) TR references.

of the references and each of the three groups, as shown below:

$$d(x_i, \mu_j) = \sqrt{(x_i - \mu_j)^T \sum_j^{-1} (x_i - \mu_j)}$$

Here, $d(x_i, \mu_j)$ is the distance between a term frequency vector (x_i) of reference i and the mean vector (μ_j) of group j (j = 1,2,3), Σ_j^{-1} is an inverse matrix of the variance-covariance matrix of group j, and T indicates the transposition of the vector. In the discriminant function, each reference was classified into the group in which d became a minimum value. To validate the discriminant function, we performed a leave-one-out cross validation (Table 3). The accuracy rates of classification were very high: 94.6% for IG and 94.7% for TR.

Table 3 Validation of reference classification by canonical discriminant analysis

Protein and group[a]		Predicted groups: Group I	Group II	Group III	Total	Accuracy (%)
IG	Group I	**111**	0	9	120	92.5
	Group II	0	**128**	11	139	92.1
	Group III	0	4	**183**	187	97.9
	Total				**422/446**	**94.6**
TR	Group I	**31**	0	3	34	91.2
	Group II	0	**33**	4	37	89.2
	Group III	0	0	**61**	61	100.0
	Total				**125/132**	**94.7**

[a]Groups I, II and III are the same as in Tables 1 and 2.

Utility

Reference and sequence search

The CIG-DB provides two search engines: (i) sequence search and (ii) reference search, which users can select on the home page.

Sequence search

Users can perform keyword searches by GI numbers, amino acid sequences, or GenBank accessions of IG/TR/epitope entries. Users can select eight search conditions, "Contains/Does not contain," "Equals/Does not equal," "Starts with/Does not start with," and "Ends with/Does not end with." The search result is shown as a table that can be sorted in ascending or descending order. Most importantly, the sequence search for IG or TR entries can be focused on cancer-related sequences by checking the filter option box. In a resultant table, GI numbers and references are linked to GenBank and PubMed online, respectively, and the reference IDs (i.e., PubMed IDs) shown are numbered as "1" (cancer therapy) or "2" (hematological tumors). The same search engine is also available for epitope entries.

Reference search

Users also can search for IG/TR entries by MeSH terms, words in the title or abstract, names of authors, journals, or publication dates, and thus narrow down research fields of interest. This feature is an advantage of the CIG-DB, because such a detailed reference search is not offered by other public immunological databases. Figure 2a shows an example of a search result using the name of an antibody medicine against breast cancer, "Herceptin." The reference search result can be reflected to the subsequent sequence search to corresponding IG/TR/epitope entries by a check box, so that users can see

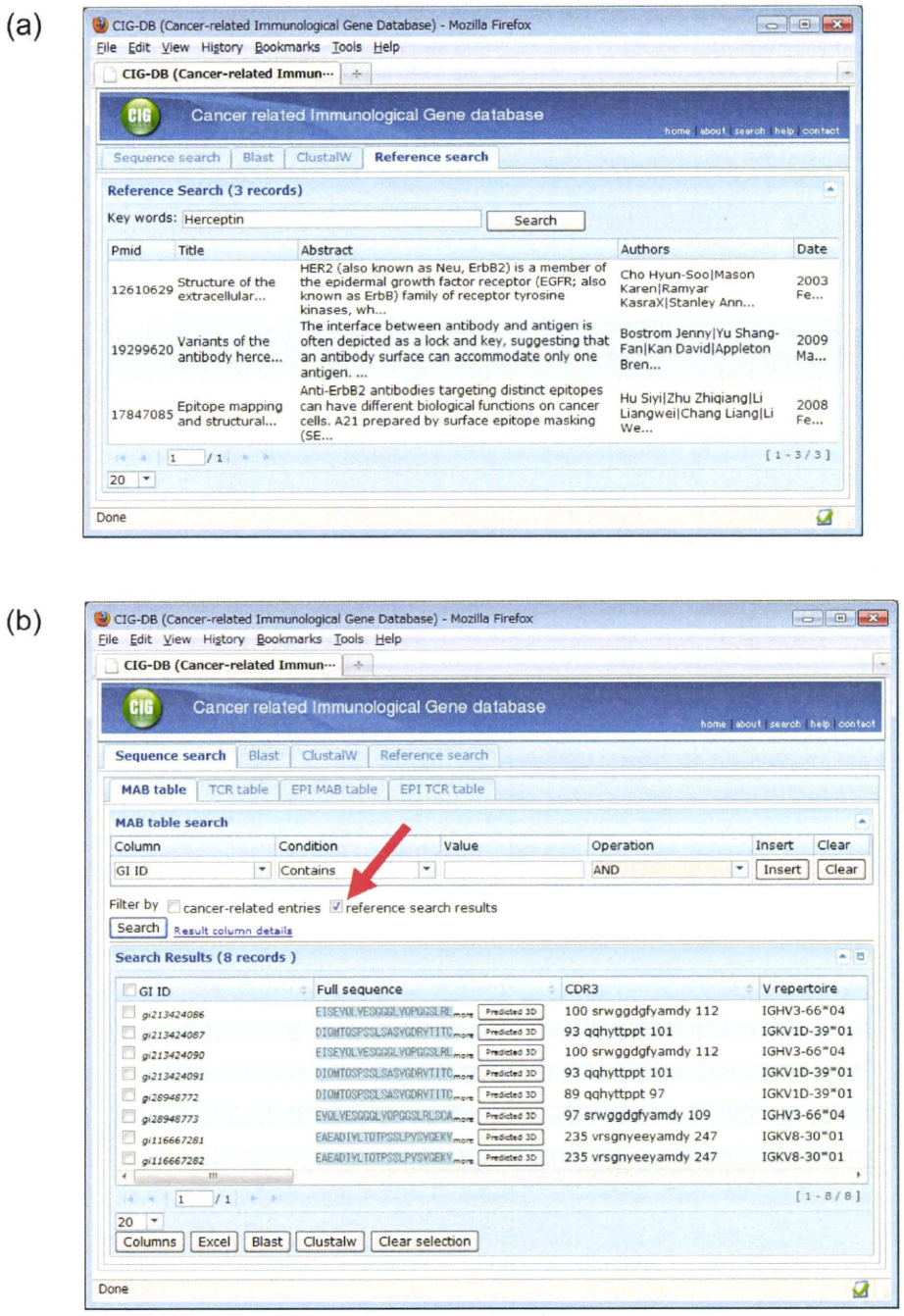

Figure 2 A search interface of the Cancer-related Immunological Gene Database (CIG-DB). (a) An example of search results where the word "Herceptin" is queried against references in the CIG-DB. (b) The subsequent sequence search result from (a). A check box for reflecting the reference search result to sequence search is indicated by a red arrow.

amino acid sequences that were reported in the matched reference (Figure 2b).

Sequence analysis tools

In clinical study, one may determine the sequences of IG or TR specific to tumor-associated antigens from a patient's B or T cells and compare these with public sequences. The CIG-DB thus provides BLAST (Ver. 2.2) and CLUSTALW [30] (Ver. 1.83) tools for sequence similarity search and alignment (Figure 3). Here users can compare the in-house sequences with cancer-related sequences in the CIG-DB, or with all public IG/TR sequences regardless of cancer studies in the proto-database. It is also possible to perform the

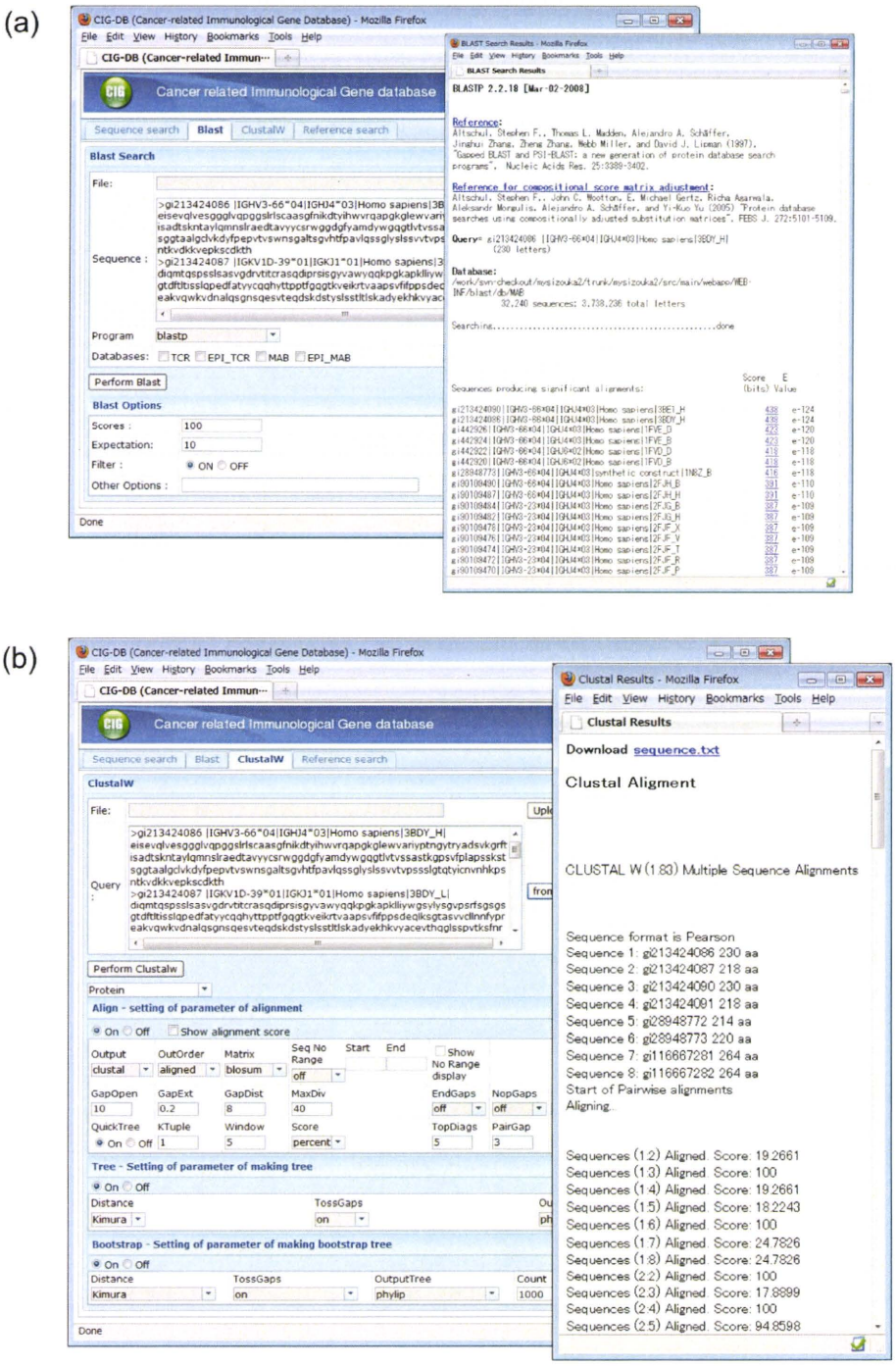

Figure 3 Sequence analysis tools of CIG-DB. (a) An example of BLAST query and result pages. (b) An example of CLUSTALW query and result pages. Both analyses were performed from the search result page in Figure 2b.

analysis using only the sequences within the CIG-DB. For example, after a reference or sequence search, the result can be subsequently utilized for such comparative analyses. The matched sequence list by BLAST and multiple alignment by CLUSTALW are downloadable in text format.

3D structure visualization and modeling

For the entries merged from the PDB, which are either for a receptor alone or a receptor-epitope complex, the tertiary structures are visualized by a JAVA applet based on Jmol [31] (Figure 4). This viewer can be customized so that interacting residues are highlighted. Particularly,

Figure 4 A 3D viewer of CIG-DB. The entries containing two words, "cancer" and "testis," were searched against the column of "Epitope name" in TR-epitope table (EPI TCR table). A 3D structure of the top hit (PDB code = 2F53, the complex between the T-cell receptor and cancer/testis antigen 1B peptide) is shown. The epitope molecule is highlighted in green by checking a box in the viewer.

in epitope tables, the epitope structure derived from the PDB is viewed as bound with the antigen receptor, and the applet allows users to investigate the interaction between CDR3 and epitope residues (atomic distance, etc.). In addition, the CIG-DB offers predicted structures for non-PDB IG/TR entries using MODELLER [32], a protein 3D modeling software based on the homology modeling method. In MODELLER, an amino acid sequence is compared against a 1D sequence database consisting of known PDB structures. PDB structures whose sequences are similar to that of the query are used as templates in 3D alignment, followed by energy optimization of aligned structures. In the CIG-DB, the templates are local data of human and mouse IG/TR proteins collected from the PDB, whose backbone structures are sufficiently conserved. For each of the non-PDB IG/TR entries, five predicted structures and their scores (objective functions) in optimization are calculated, and users can download these models in PDB format for further refinement by structure calculation (e.g., molecular dynamics or docking simulation).

Database implementation and update

The CIG-DB is a Java web application developed using an open source Ajax framework, ZK, which uses MySQL as a database backend. As a high performance search engine, this database is equipped with Apache Lucene. The application runs on a servlet container, Apache Tomcat. An update of the basic contents is semi-automatic, achieved by Perl and Shell scripts. In the future, the reference classification script by R programs will be integrated into the automation.

Discussion

Immunotherapy is now an effective treatment for cancer, where the information on cancer-antigen-specific IG or TR as well as the epitopes, is essential for its use. In particular, genetic engineering, such as the improvement of antibody sequences as molecular target drugs or the design of epitope peptides for cancer vaccines, are of great interest to clinicians and pharmacologists. For such studies, a sequence-based database for cancer immunological genes has potential utility. In addition, sequence comparison tools and structural data are useful for genetic engineering, combined with library screening methods in a laboratory [33,34]. Our database, the CIG-DB, thus may meet the needs of researchers involved in such cancer studies. Moreover, this database is also designed as a reference-based platform, equipped with a search engine by MeSH terms, title, and abstract words (Figure 2a). In previous public databases, one could find the immunological gene sequences, but the references were not always related to cancer, or one could search for cancer-related papers, but the sequences frequently were undetermined. Our database avoids such inconveniences and users can easily obtain the cancer-related IG/TR sequences that are available. For comparison, IMGT also provides with the data sheets on therapeutic monoclonal antibodies related to oncology http://www.imgt.org/mAb-DB/query, but does not yet encompass the information about cancer vaccines and TRs specific to tumor-associated antigens. The IEDB is an exhaustive epitope resource and covers a wide range of experimental details, but the search option is not optimized to find cancer-related protein/peptide sequences.

Considering their application in cancer studies, IG and TR proteins are derived from two mammals, the human and the mouse. In this study, it was found that such cancer-related IG/TR proteins can be classified into two groups, namely, cancer therapy (Group I) and hematological tumors (Group II). One more group (Group III) of unrelated references includes papers mainly involving experiments about hybridoma using "myeloma" or those about irrelevant "tumors," which are therefore wrongly selected by keyword matching. The classification of cancer-related IG and TR is thus very important for ensuring a high quality of the CIG-DB. Maintenance of the current version of the CIG-DB is semi-automated from initially obtaining the sequence data to presenting the final tables on the graphical user interface. Manual operation is only necessary for the very classification of IG and TR annotation entries into the two groups by

checking their references, but this method would be a troublesome task for future updates. To reduce this burden, we prepared a computational classification program inside the database. As a basis for this step, PCA results suggest that the two groups can be discriminated from each other by the term frequencies in MEDLINE texts (Figure 1). Although Group III overlaps somewhat with Group II in the case of IG references, the distributions can be statistically discriminated from each other ($P < 0.001$, multivariate analysis of variance). In this study, we calculated a canonical discriminant function using a set of references that were classified manually in the current version of the database as a training dataset. We evaluated the function by leave-one-out cross validation and obtained high accuracy rates (Table 3), strongly suggesting that our method can be applicable for an automated update. For the next update, this procedure may possibly be integrated into the maintenance programs of the CIG-DB. Alternatively, we will perform a manual classification and calculate the discriminant function again over a few more updates, further improving the accuracy of the database.

This database provides structural data of antigen receptors. Particularly, for all non-PDB IG/TR entries, the five predicted structural models are available in PDB format, which allows users to study the interaction between IG/TR and the epitope using molecular dynamics or docking simulation. This is an advantage of our database over other public databases, such as IMGT and the IEDB which provide only known structures. It should be noted that the combinatorial study using these tools and IG/TR data in Group I are efficient for cancer immunotherapy, in which genetic design and engineering are performed for developing novel antibody medicines, tumor-specific TR, and peptide vaccines with potent anticancer effects. Recently, new technologies have allowed rapid cloning of effective antibodies to specific antigens [35,36]. It is thus likely that a large number of antibodies and TR for tumor-associate antigens will be open for genetic improvement in the near future. Since the CIG-DB is a semi-automated database, the latest information will be quickly reflected there. We believe that accumulation of IG/TR/epitope data will enhance the usefulness of this database in clinical cancer studies.

Conclusions

The CIG-DB is designed to serve as a bridge between immunological gene data and cancer studies, presenting annotations of IG, TR, and their epitopes. In its current version, the database has 2,081 cancer-related human and murine receptor entries (1,605 for IG and 476 for TR), and 920 entries for epitopes bound with receptors from a variety of organisms. Regarding IG and TR proteins, this database further provides a helpful guide to two detailed groups; one for cancer therapy and the other for hematological tumors. For the next update, we have developed a powerful method to automatically classify cancer-related entries into these two groups. The high precision (~95% accuracy) shown in validation assessments is promising for the efficient performance of the database. Moreover, the CIG-DB is equipped with sequence- and reference-search engines and analysis tools, the results of which can be utilized for advanced studies. In particular, the database will play important roles in cancer immunotherapy, by integrating the accumulating patient-specific IG/TR/epitope sequence data by novel cloning technologies.

Availability and requirements

The CIG-DB is available without charge or registration at http://www.scchr-cigdb.jp/. We have confirmed that the web site can be viewed by Internet Explorer 7 or later, Safari 3, and Firefox 3. The Java Runtime Environment (JRE 1.5 or higher) is required for displaying 3D structures by Jmol applet.

Acknowledgements

We thank A. U. Umagiliya of Bioinformatics Institute of Global Good Inc. (BiGG) for technical supports in database construction, and A. Iizuka, K. Ozawa and M. Komiyama for database test and helpful comments. We are also grateful to T. Makino for preparing prototype scripts. This work was supported in part by a grant in Cooperation of Innovative Technology and Advanced Research in an Evolutional Area (CITY AREA) from the Ministry of Education, Culture, Sports, Science and Technology, Japan (MEXT).

Author details

[1]Immunotherapy Division, Shizuoka Cancer Center Research Institute, 1007 Shimonagakubo, Nagaizumi-cho, Sunto-gun, Shizuoka, 411-8777, Japan. [2]Department of Clinical Pharmacology, Tokai University School of Medicine, 143 Shimokasuya, Isehara, Kanagawa, 259-1193, Japan. [3]Bioinformatics Institute for Global Good Inc., Kitashinagawa 3-6-9, Shinagawa-ku, Tokyo, 140-0001, Japan. [4]Center for Information Biology and DNA Data Bank of Japan, National Institute of Genetics, Yata 1111, Mishima, Shizuoka, 411-8540, Japan. [5]*Current Address:* National Research Institute of Fisheries Science, Fisheries Research Agency, 2-12-4 Fukuura, Kanazawa, Yokohama, Kanagawa, 236-8648, Japan.

Authors' contributions

YN designed the database, developed the core scripts and wrote the manuscript. TK and YA conceived the study. MF developed the Web interface. TG participated in coordinating the study and helped to draft the manuscript. All authors read and approved the final manuscript.

Received: 22 April 2010 Accepted: 27 July 2010 Published: 27 July 2010

References

1. Abbas AK, Lichtman AH, Pillai S: *Cellular and Molecular Immunology* Philadelphia: Saunders, 6 2007.
2. Tonegawa S: **Somatic generation of antibody diversity.** *Nature* 1983, **302(5909)**:575-581.
3. Chen C, Stenzel-Poore MP, Rittenberg MB: **Natural auto- and polyreactive antibodies differing from antigen-induced antibodies in the H chain CDR3.** *J Immunol* 1991, **147(7)**:2359-2367.
4. Davies DR, Cohen GH: **Interactions of protein antigens with antibodies.** *Proc Natl Acad Sci USA* 1996, **93(1)**:7-12.

5. Jorgensen JL, Esser U, Fazekas de St Groth B, Reay PA, Davis MM: **Mapping T-cell receptor-peptide contacts by variant peptide immunization of single-chain transgenics.** *Nature* 1992, **355(6357)**:224-230.
6. Forgber M, Trefzer U, Sterry W, Walden P: **Proteome serological determination of tumor-associated antigens in melanoma.** *PLoS One* 2009, **4(4)**:e5199.
7. Tan EM, Zhang J: **Autoantibodies to tumor-associated antigens: reporters from the immune system.** *Immunol Rev* 2008, **222**:328-340.
8. Nagorsen D, Scheibenbogen C, Marincola FM, Letsch A, Keilholz U: **Natural T cell immunity against cancer.** *Clin Cancer Res* 2003, **9(12)**:4296-4303.
9. Nowell PC: **Chromosomal approaches to hematopoietic oncogenesis.** *Stem Cells* 1993, **11(1)**:9-19.
10. Waldmann TA: **Immunotherapy: past, present and future.** *Nat Med* 2003, **9(3)**:269-277.
11. Finn OJ: **Tumor immunology top 10 list.** *Immunol Rev* 2008, **222**:5-8.
12. Liu XY, Pop LM, Vitetta ES: **Engineering therapeutic monoclonal antibodies.** *Immunol Rev* 2008, **222**:9-27.
13. Kubota T, Niwa R, Satoh M, Akinaga S, Shitara K, Hanai N: **Engineered therapeutic antibodies with improved effector functions.** *Cancer Sci* 2009, **100(9)**:1566-1572.
14. Ross JS, Fletcher JA: **The HER-2/neu oncogene in breast cancer: prognostic factor, predictive factor, and target for therapy.** *Stem Cells* 1998, **16(6)**:413-428.
15. Itoh K, Yamada A, Mine T, Noguchi M: **Recent advances in cancer vaccines: an overview.** *Jpn J Clin Oncol* 2009, **39(2)**:73-80.
16. Varela-Rohena A, Carpenito C, Perez EE, Richardson M, Parry RV, Milone M, Scholler J, Hao X, Mexas A, Carroll RG, *et al*: **Genetic engineering of T cells for adoptive immunotherapy.** *Immunol Res* 2008, **42(1-3)**:166-181.
17. Stewart-Jones G, Wadle A, Hombach A, Shenderov E, Held G, Fischer E, Kleber S, Nuber N, Stenner-Liewen F, Bauer S, *et al*: **Rational development of high-affinity T-cell receptor-like antibodies.** *Proc Natl Acad Sci USA* 2009, **106(14)**:5784-5788.
18. Robbins PF, Li YF, El-Gamil M, Zhao Y, Wargo JA, Zheng Z, Xu H, Morgan RA, Feldman SA, Johnson LA, *et al*: **Single and dual amino acid substitutions in TCR CDRs can enhance antigen-specific T cell functions.** *J Immunol* 2008, **180(9)**:6116-6131.
19. Parkhurst MR, Joo J, Riley JP, Yu Z, Li Y, Robbins PF, Rosenberg SA: **Characterization of genetically modified T-cell receptors that recognize the CEA:691-699 peptide in the context of HLA-A2.1 on human colorectal cancer cells.** *Clin Cancer Res* 2009, **15(1)**:169-180.
20. Schoonbroodt S, Steukers M, Viswanathan M, Frans N, Timmermans M, Wehnert A, Nguyen M, Ladner RC, Hoet RM: **Engineering antibody heavy chain CDR3 to create a phage display Fab library rich in antibodies that bind charged carbohydrates.** *J Immunol* 2008, **181(9)**:6213-6221.
21. Yoon SO, Lee TS, Kim SJ, Jang MH, Kang YJ, Park JH, Kim KS, Lee HS, Ryu CJ, Gonzales NR, *et al*: **Construction, affinity maturation, and biological characterization of an anti-tumor-associated glycoprotein-72 humanized antibody.** *J Biol Chem* 2006, **281(11)**:6985-6992.
22. Lefranc MP, Giudicelli V, Kaas Q, Duprat E, Jabado-Michaloud J, Scaviner D, Ginestoux C, Clement O, Chaume D, Lefranc G: **IMGT, the international ImMunoGeneTics information system.** *Nucleic Acids Res* 2005, , **33 Database:** D593-597.
23. Vita R, Zarebski L, Greenbaum JA, Emami H, Hoof I, Salimi N, Damle R, Sette A, Peters B: **The Immune Epitope Database 2.0.** *Nucleic Acids Res* 2009.
24. Lefranc M-P, Lefranc G: **The Immunoglobulin FactsBook.** Academic Press, London, UK 2001.
25. Giudicelli V, Chaume D, Lefranc MP: **IMGT/GENE-DB: a comprehensive database for human and mouse immunoglobulin and T cell receptor genes.** *Nucleic Acids Res* 2005, , **33 Database:** D256-261.
26. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucleic Acids Res* 1997, **25(17)**:3389-3402.
27. Berman HM, Battistuz T, Bhat TN, Bluhm WF, Bourne PE, Burkhardt K, Feng Z, Gilliland GL, Iype L, Jain S, *et al*: **The Protein Data Bank.** *Acta Crystallogr D Biol Crystallogr* 2002, **58(Pt 6 No 1)**:899-907.
28. Saha S, Bhasin M, Raghava GP: **Bcipep: a database of B-cell epitopes.** *BMC Genomics* 2005, **6(1)**:79.
29. Korber BTM, Brander C, Haynes BF, Koup R, Moore JP, Walker BD, Watkins DI, (eds): **HIV Molecular Immunology.** Los Alamos National Laboratory 2006.
30. Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R, *et al*: **Clustal W and Clustal X version 2.0.** *Bioinformatics* 2007, **23(21)**:2947-2948.
31. **Jmol: an open-source Java viewer for chemical structures in 3D.** [http://jmol.sourceforge.net/].
32. Sali A, Blundell TL: **Comparative protein modelling by satisfaction of spatial restraints.** *J Mol Biol* 1993, **234(3)**:779-815.
33. Dona MG, Giorgi C, Accardi L: **Characterization of antibodies in single-chain format against the E7 oncoprotein of the human papillomavirus type 16 and their improvement by mutagenesis.** *BMC Cancer* 2007, **7**:25.
34. Ni M, Yu B, Huang Y, Tang Z, Lei P, Shen X, Xin W, Zhu H, Shen G: **Homology modelling and bivalent single-chain Fv construction of anti-HepG2 single-chain immunoglobulin Fv fragments from a phage display library.** *J Biosci* 2008, **33(5)**:691-697.
35. Jin A, Ozawa T, Tajiri K, Obata T, Kondo S, Kinoshita K, Kadowaki S, Takahashi K, Sugiyama T, Kishi H, *et al*: **A rapid and efficient single-cell manipulation method for screening antigen-specific antibody-secreting cells from human peripheral blood.** *Nat Med* 2009, **15(9)**:1088-1092.
36. Wrammert J, Smith K, Miller J, Langley WA, Kokko K, Larsen C, Zheng NY, Mays I, Garman L, Helms C, *et al*: **Rapid cloning of high-affinity human monoclonal antibodies against influenza virus.** *Nature* 2008, **453(7195)**:667-671.

doi:10.1186/1471-2105-11-398
Cite this article as: Nakamura *et al.*: CIG-DB: the database for human or mouse immunoglobulin and T cell receptor genes available for cancer studies. *BMC Bioinformatics* 2010 **11**:398.

Funct Integr Genomics (2010) 10:307–319
DOI 10.1007/s10142-010-0168-1

REVIEW

The transcript repeat element: the human Alu sequence as a component of gene networks influencing cancer

Paula Moolhuijzen · Jerzy K. Kulski · David S. Dunn · David Schibeci · Roberto Barrero · Takashi Gojobori · Matthew Bellgard

Received: 2 December 2009 / Revised: 18 February 2010 / Accepted: 19 February 2010 / Published online: 15 April 2010

Abstract A small percentage (3%) of the 1.3 million copies of Alu sequences in the human genome is expressed individually or as part of various gene transcripts with potential regulatory and pathophysiological importance. In order to better understand the role of repetitive elements within transcripts, this review focuses on Alu-containing transcripts of normal and cancerous tissue in a transcriptome-wide survey of the H-Invitational human transcript database on 106,825 tissue-derived transcripts expressed at 29,979 loci. The Alu elements in transcripts of cancerous tissues are significantly underrepresented in comparison to those in normal tissues. In this review, we propose a model for Alu-mediated siRNA down-regulation of Alu-containing transcripts in cancer tissues. In cancer or other rapidly dividing tissues, hypomethylation of repeat element regions triggers the expression of transposon elements including Alu, which can potentially form double-stranded RNA molecules for use as templates to generate Alu-derived siRNAs (Alu-siRNAs). The generated Alu-siRNAs target endogenous messenger RNAs harbouring sequence similarity to Alu elements. This model correlates with the observation that there is substantial under-representation of Alu-containing mRNAs in cancer cells. This new perspective of gene regulation in disease conditions can provide a basis for starting to account for changes in complex gene network in cancer.

Keywords Bioinformatics · Comparative genomics · Disease · Gene control

Paula Moolhuijzen and Jerzy K. Kulski contributed equally to this work.

Electronic supplementary material The online version of this article (doi:10.1007/s10142-010-0168-1) contains supplementary material, which is available to authorized users.

P. Moolhuijzen · J. K. Kulski · D. S. Dunn · D. Schibeci · R. Barrero · M. Bellgard (✉)
Centre for Comparative Genomics, School for Information Technology, Murdoch University,
Murdoch, WA, Australia
e-mail: mbellgard@ccg.murdoch.edu.au

P. Moolhuijzen
e-mail: pmoolhuijzen@ccg.murdoch.edu.au

J. K. Kulski
e-mail: kulski@me.com

D. S. Dunn
e-mail: ddunn@murdoch.edu

D. Schibeci
e-mail: dschibeci@ccg.murdoch.edu.au

R. Barrero
e-mail: rbarrero@ccg.murdoch.edu.au

J. K. Kulski
Division of Molecular Life Science, Department of Genetic Information, School of Medicine, Tokai University,
Isehara, Kanagawa, Japan

J. K. Kulski
Centre for Forensic Science, The University of Western Australia,
Nedlands, WA, Australia

T. Gojobori
Center for Information Biology and DNA Data Bank of Japan
National Institutes of Genetics,
Mishima, Shizuoka, Japan
e-mail: tgojobor@genes.nig.ac.jp

Introduction

Transposable elements (TEs) contribute to about half the content of the human genome and have been found in abundance in gene sequences and in a significant portion of mature mRNAs. Most of the TEs in the human genome are retroelements, such as the long terminal repeat (LTR), short interspersed nuclear element (SINE) Alu, long interspersed nuclear element (LINE) L1 and mammalian interspersed repeat sequences that have spread throughout the genome by retrotransposition. Retrotransposons copy themselves by transcription to RNA, followed by reverse transcription to DNA and then insertion at new positions in the genome. Retrotransposition and recombination of TEs can contribute to human molecular evolution and inherited disease, organise the genome into active and inactive regions, separate domains and functional regions within a chromatin domain, suppress transcriptional noise and regulate transcript stability (Schulz et al. 2006).

Transposable elements influence gene expression

TEs can influence gene expression by a number of methods previously reviewed (Feschotte 2008; Jamalkandi and Masoudi-Nejad 2009; Kim 2005). At the transcriptional level, a TE inserted upstream of a gene may (1) insert promoter sequences and introduce an alternative transcription start site, (2) disrupt existing *cis*-regulatory element(s) or (3) introduce a new *cis*-element such as a transcription factor binding site. TE inserted within an intron may drive antisense transcription and potentially interfere with sense transcription, and TE may serve as a nucleation centre for the formation of heterochromatin potentially silencing the transcription of adjacent gene(s) (Feschotte 2008).

At the post-transcriptional level, TE inserted in the three prime (3′) untranslated region (UTR) of a gene may introduce an alternative polyadenylation site, a binding site for a microRNA (miRNA) or for RNA-binding proteins, whereas TE inserted within an intron can interfere with the normal splicing pattern of a pre-mRNA, provoking various forms of alternative splicing (intron retention and exon skipping; Feschotte 2008). Intronic Alu elements inserted in opposite orientation can undergo base-pairing and affect the splicing patterns of a downstream exon, shifting it from constitutive to alternative coding (Lev-Maor et al. 2008); TE inserted within an intron containing cryptic splice sites may be incorporated ('exonised') as an alternative exon. This may result in the translation of a new protein isoform or in the destabilisation or degradation of the mRNA via the nonsense mediated decay pathway, especially if the exonised TE introduces a premature stop codon (Feschotte 2008). The RNA editing of Alu sequences (Kim et al. 2004; Levanon et al. 2004; Maas et al. 2003; Paz et al. 2007) also introduces a new level of variation for the biological effects of these elements (Supplementary document 1).

Small RNA: repeat derived sequences as regulatory networks

Feschotte (2008) reviewed that TEs movement and accumulation provide abundant material from which *cis*-regulatory elements emerge de novo through the introduction of a single or a few point mutations (Britten 1996; Feschotte 2008; Hambor et al. 1993; Medstrand et al. 2005; Zhou et al. 2002), and the dispersal of expanding TE families throughout genomes potentially allows the same regulatory motif(s) to be recruited at many chromosomal locations, drawing multiple genes into the same regulatory network (Britten and Davidson 1969, 1971; Feschotte 2008; Johnson et al. 2006; Peaston et al. 2004). Novikova (2009) suggested that chromodomains could be responsible for the targeted integration of LTR retrotransposons, which should be favourable for mobile elements allowing them to avoid negative selection arising from insertion into coding regions.

Feschotte (2008) proposed that the de novo assembly of a small RNA network from a TE family combines the idea of TE-host gene co-transcription with the origin of a small RNA precursor containing a TE of the same family, and that the precursor may arise by transcription and intramolecular folding of a TE with a nearly perfect palindromic structure (e.g. MITEs). The resulting double-stranded RNA may then be processed into a mature miRNA, and the resulting TE-derived miRNA can then pair with complementary TE sequences embedded within the 3′ UTR of co-transcribed mRNAs (Feschotte 2008).

As small non-protein-coding RNAs (ncRNAs) are important components in the regulation of eukaryotic gene expression (Mattick 2007), several classes of small regulatory RNA, including miRNAs, small interfering RNAs (siRNA), repeat-associated small interfering RNAs (rasiRNAs) and piwi-interacting RNAs (piRNAs), use partially overlapping pathways akin to RNA interference (RNAi) to silence gene expression, via degradation or translation inhibition of mRNAs containing complementary sites (Feschotte 2008). A fully integrated small RNA pathway connecting ncRNA entities such as miRNAs, siRNAs, trans-acting small interfering RNA (ta-siRNAs) and natural antisense transcripts pathway involving naturally occurring siRNAs (nat-siRNAs) have been recon-

structed within *Arabidopsis thaliana* (Jamalkandi and Masoudi-Nejad 2009). Transcriptional gene silencing of elongation factor 1 alpha has been shown in human tissue culture cells to inhibit mRNA transcription by promoter-directed siRNA (Morris et al. 2004).

Small RNAs mediate gene silencing through at least four different mechanisms: (1) endonucleolytic cleavage of the cognate mRNAs, (2) translational repression, (3) transcriptional repression through the modification of DNA and/or histone and (4) DNA elimination through the modification of histone (Kim 2005).

Feschotte (2008) proposed that post-transcriptional regulation of multiple genes through the recognition of shared *cis*-elements by a single small RNA species is similar to the logic of transcriptional regulation by transcription factors (Chen and Rajewsky 2007) which he noted was supported by the fact that some small RNAs are able to mediate homology-dependent transcriptional silencing and participate in the nucleation of heterochromatin (Grewal and Jia 2007; Slotkin and Martienssen 2007).

The relationship of piRNAs, siRNAs and rasiRNAs to TEs has been reported (Jamalkandi and Masoudi-Nejad 2009), and the proposed function of these small RNAs is to silence the expression of invasive DNA viruses and control the replication of TEs (Aravin et al. 2007; Slotkin and Martienssen 2007). Several mammalian miRNA precursors have been found to contain or be derived from TE sequences (Piriyapongsa et al. 2007; Smalheiser and Torvik 2005), and a substantial number of predicted miRNA targets map within members of the same TE families (Piriyapongsa et al. 2007; Smalheiser and Torvik 2005, 2006), again pointing to a model whereby large sets of *cis*-regulatory sequences have been seeded by transposition (Feschotte 2008). Approximately 12% experimentally characterised human miRNA genes have originated from TEs (Piriyapongsa et al. 2007), a level lower than TEs in the transcriptome (Feschotte 2008).

Paramutation is the ability of specific homologous DNA sequences to communicate in trans to establish meiotically heritable expression/epigenetic states. Intriguingly, newly silenced sequences continue to issue instructions to naive alleles in subsequent generations (Chandler and Alleman 2008; Chandler and Stam 2004). Paramutation interactions have been described in several kingdoms, including human. Chandler and Stam (2004) hypothesised that multiple mechanisms underlie the diverse phenomena and suggest two non-mutually exclusive models; pairing interactions between specific chromatin complexes and trans-RNA-based communication.

Chandler and Stam (2004) reviewed in maize, systems for which the regulatory sequences required for paramutation have been identified, genes *b1* and *p1* that encode transcription factors that activate the biosynthesis of flavonoid pigments in plant. Stam et al. (2002) identified a seven tandem repeat sequence 853-bp long upstream of the *b1* transcription site that was highly methylated at *B-I* allele (dark purple) relative to *B′* (light purple). Conversion of *B-I* to *B′* was specifically associated with a reduced methylation of the tandem repeat region (Stam et al. 2002). An RNA-mediated trans-induction of chromatin mechanism was proposed to explain the role of siRNA.

The dsRNA that is formed by transcription from the two strands of the repeated DNA is a target for Dicer, which produces siRNA. The siRNA is then postulated to mediate chromatin changes, which in turn alters the expression of the adjacent gene. Possible mechanisms include, but are not limited to, RNA-directed DNA methylation and RNA-directed histone modification (Chandler and Stam 2004).

The dsRNAs generated from the repeats by sense and rare antisense transcription produce siRNAs that could be efficiently amplified from tandem array transcripts by RNA-dependent RNA polymerase (RdRP) and Dicer; the siRNA primers can prime from an upstream repeat, replenishing siRNA for the whole repeat sequence. RdRP activity, which synthesises complementary strand RNA using siRNA primers, results in increased amounts of siRNAs throughout the repeats. The production of dsRNA that trigger RNAi in turn can result in degradation of homologous mRNA, altered chromatin states associated with DNA methylation or, potentially, inhibition of translation (Chandler and Stam 2004).

Expression of Alu repetitive elements

Of the various retroelement families, the primate-specific Alu, so-called because it contains and was identified by its recognition site for the restriction endonuclease AluI (Houck et al. 1979), is the largest family of SINEs with an estimated 1.3 million copies contributing to 10% of the human genome (Lander et al. 2001; Price et al. 2004). Alu elements are dimeric sequences with a characteristic length of 300 bp that probably originated from a gene encoding the 7SL RNA (Ullu and Tschudi 1984) and then developed into dimeric sequences via the Alu monomeric forms FAM, FRAM and FLAM (Quentin 1992a, b). The full-length 300-bp Alu element has evolved in primates into a number of different dimeric families and subfamilies that are distinguished from each by diagnostic markers within their sequences as well as by their evolutionary history or phylogeny (Kapitonov and Jurka 1996). Essentially, the three main Alu families or classes categorised as Alu-J, Alu-S and Alu-Y were estimated by sequence divergence to

have evolved specifically in primates about 65–80 million years ago (mya), 30–50 mya, and 15–30 mya, respectively (Kapitonov and Jurka 1996). Each of the three Alu families has additional subfamily members that can be identified by sequence alignments with consensus subfamilies using the computer program Repeatmasker3 (Smit et al. 1996–2004), Censor (Jurka et al. 1996) or other programs (Price et al. 2004). The Alu family found most commonly in transcripts and the human genome is Alu-S (Price et al. 2004). Studies have surveyed the numbers of transcript (Dagan et al. 2004; Makalowski 2000), but the overall proportion of Alu-containing transcript (Alu-transcript) has remained essentially the same (Yulug et al. 1995).

Fragmented and/or full-length Alu elements have been found in the coding regions of mRNA and may be beneficial, neutral or deleterious to the function of the gene and its transcripts. Alu RNA levels have been reported to increase in response to cell stress (Chu et al. 1998; Hagan et al. 2003). Alu sequences are known to regulate gene expression and translation at the transcriptional and post-transcriptional levels (Hasler and Strub 2006), modulate cellular growth, differentiation and tumour suppression (Vila et al. 2003) and function in exonisation (Sorek et al. 2002). For example, Alu elements are a source of adenine and uracil-rich elements at the 3′ UTR that may contribute to the stabilisation or degradation of mRNA (An et al. 2004). Approximately 5% of alternately spliced internal exons in the human genome was found to have an Alu sequence (Sorek et al. 2002), with most Alu exons alternately spliced and with only a segment of the Alu sequence contributing to the new open reading frame (ORF; Krull et al. 2005).

Earlier studies suggested that Alu transcription is regulated by epigenetic mechanisms such as DNA methylation and histone modification at Alu repeats (Liu et al. 1994; Kondo and Issa 2003). Saito et al. (2009) found in gastric cancer that chromatin remodelling at Alu-S repeats by DNA demethylation and HDAC inhibition can activate expression of Alu-associated miRNAs, which can down-regulate target oncogenes in human (Saito et al. 2009).

Human Alu repeat sequence as a component of gene networks

In order to better understand the TE content of transcripts and their possible functions in human, we surveyed and analysed two of the public transcript databases: the H-Invitational (H-Inv; Japan; http://www.h-invitational.jp/) and NCBI (USA; http://www.ncbi.nlm.nih.gov/) human transcript datasets that are the largest and most complete collections of human transcript sequence information currently available.

Differentiating between the cancerous and normal tissue information within the integrated H-Inv database

The H-Inv human full-length transcript annotation project (http://www.h-invitational.jp/) provides transcript accessions representing 35,000+ loci in the H-Inv database (H-InvDB). This integrated database has been highly curated by experts to provide high-quality manual and computational annotation of human genes and transcripts and information about gene structures, alternative splicing isoforms and non-coding functional RNAs (Yamasaki et al. 2005).

To ensure that the H-Inv and NCBI human transcript datasets were of sufficient quality for investigation, additional filters were applied as follows: (1) Transcripts derived from culture cells were removed, and only those found in tissues retained. (2) Association of transcripts with cancerous tissues was determined by evaluating a match to the following 20 terms: adenocarcinoma, astrocytoma, carcinoma, choriocarcinoma, glioma, glioblastoma, gliosarcoma, hepatoma, leukaemia, lymphoma, melanoma, melanotic, neuroblastoma, neuroepithelioma, papilloma, phaeochromocytoma, retinoblastoma, rhabdomyosarcoma, teratocarcinoma and tumour. (3) Tissue information was then categorised by organ type. (4) The dataset sizes for normal or cancerous tissues were examined, and underrepresented tissues were excluded from further analysis. (5) The transcript length qualities were evaluated in order to assess the Alu positional insertion within the transcript. Transcripts can be divided into three regions, five prime UTR (5′ UTR), coding sequence (CDS) and three prime UTR (3′ UTR), and Alu content was represented based on quantified base-pair percentage and the fraction of total transcript numbers. To further assess the quality of the transcripts between the two categories of normal and cancerous tissues, housekeeping genes can be investigated to determine size bias to the 3′ UTR regions from normal and cancerous tissues. The average length of the 3′ UTR sequence for the 149 expressed genes was 558.3 nucleotides for transcripts derived from cancerous tissues and 957.9 nucleotides for transcripts derived from normal tissues, and the t test significant ($p<0.05$) at $p=0.045$. Since the statistical difference between the normal and cancerous tissues for the length of their 3′ UTR was relatively borderline, it was assumed that the transcripts were of sufficient quality in both cancerous and normal tissues to undertake the following further comparisons.

The development and public access to primary and secondary databases based on sequence information gives an opportunity to investigate the distribution of Alu in transcripts. Previous analyses of Alu-transcripts have resulted in the creation of databases that highlight Alu positions within transcript (Makalowski 2000). In order to

better understand the possible regulatory functions of Alu within RNA, surveys of full-length transcript from a large array of normal and cancerous tissues were used to compile transcripts encompassing all the nucleotide sequences from the CAP site to the poly (A) addition site or at least the entire CDS of a protein. The H-InvDB provides 167,992 full-length and partial transcripts encoded by 35,005 human gene loci (Imanishi et al. 2004). The loci gene transcript structures and alternative splicing forms referred to here as isoforms can be found at http://www.h-invitational.jp/.

All sequence annotation on tissue and cell type, locus Entrez gene identifier and the GenBank CDS start and end positions can be extracted from NCBI GenBank (http://www.ncbi.nlm.nih.gov/sites/entrez?db=nucleotide) using Bioperl tools (Stajich et al. 2002) and H-InvDB downloadable text file 'FCFUN' (ftp://hinv.ddbj.nig.ac.jp/). The H-InvDB functional annotation categories however are also conveniently available as follows: category I, identical to a known human protein with greater than or equal to 98% identity and 100% coverage; category II, similar to known protein in another species with greater than or equal to 50% identity; category III, containing an InterPro (Apweiler et al. 2001) protein domain; category IV, conserved hypothetical protein, greater than or equal to 50% identity to a known hypothetical protein; category V: hypothetical protein with an ORF length greater than or equal to 80 amino acids, no pseudogene overlaps; category VI, hypothetical short protein, with an ORF length less than 80 amino acids; category VII, pseudogene candidates.

For the total 167,992 H-Inv transcript sequences, 107,001 sequences are derived from tissue that can be subdivided into gene loci and the number of Alu-transcript in normal and cancerous tissues. Table 1 shows the relative numbers and percentages for the distribution of the gene or genomic loci-containing transcripts with and without Alu sequences. It was also determined if these loci transcripts were either unique or common to both the cancerous and normal tissues. For the total 106,825 transcripts, 33,918 transcript represented by 13,798 loci were assigned to a cancerous tissue trait, and 72,907 transcript represented by 26,677 loci were assigned to a normal tissue trait. The number of Alu-transcript totaled 17,861 (17%) of the 106,825 transcripts derived from tissues and was represented by 13,240 loci or 44% of the total gene loci. This set of Alu-transcript is represented by 14,191 normal-sourced tissue sequences (10,648 loci) and 3,670 cancer-sourced tissue sequences (2,592 loci), represented by 19% and 11% of normal and cancerous sourced tissue sequences, respectively. For the represented 29,979 loci, 3,302 loci (11%) produce transcript derived only from cancerous tissues, 16,181 loci (54%) produced transcript only from normal tissues and 10,496 (35%) of loci contained transcript produced from both cancerous and normal tissues.

The fraction and relative tissue distribution of Alu-transcript to transcript within the cancerous or the normal tissue group is shown in Fig. 1. Of the 25 tissues examined, 22 had a statistically significant proportional difference of Alu-transcript to transcript in the normal to the cancerous tissues as determined by Fisher's exact two-sided tests ($P<0.001$ for 16 tissues and $P<0.01$ for six tissues). The four cancerous tissues with a significantly ($P<0.001$) greater proportion of Alu-transcript to transcript in them than in the normal tissues were liver, oral cavity, ovary and placenta. Of these four tissues, the greatest relative difference between the normal and cancerous tissue for Alu-transcript to transcript was in the ovary. Oesophagus, pancreas and skin showed no significant ($p>0.05$) difference between normal and cancerous tissues. Of the cancerous tissues, the greatest proportion of Alu-transcript to transcript was in the oral tissues, whereas of the normal tissues, the greatest proportion of Alu-transcript to transcript was in the rectal tissues.

Alu-transcript functions

The functional categories of the Alu-transcript collected from the H-Inv human transcript database were analysed by Gene Ontology (GO), which classifies the known functions of gene products into at least 5,175 categories according to biological processes, cellular components and molecular functions (Ashburner et al. 2000). GO analysis was performed for transcripts with an NCBI gene identifier and CDS to determine the different metabolic and regulatory pathways that are possibly associated with the Alu-transcript in normal and cancerous tissues. Over 8,000 Alu-transcripts from the H-InvDB were classified into nearly 3,000 of the GO

Table 1 The percentage and number of unique and overlapping loci groups for loci-containing transcript and Alu-containing transcript (Alu-transcript) sequences in normal and cancerous tissues

	No. of loci	Normal only	Cancer only	Both normal and cancer	No. of transcript	No. of Alu-transcript	%	No. of loci with Alu-transcript	Percentage of loci with Alu-transcript
Normal condition	26,677	16,181	–	10,496	72,907	14,191	19	10,648	39.91
Cancer condition	13,798	–	3,302		33,918	3,670	11	2,592	18.79
Total		29,979			106,825	17,861	17	13,240	44.16

The number and percentage transcript and Alu-transcript derived from normal and cancerous tissues