Table I. *The dn/ds ratios in the known CTL epitopes restricted to HLA-A*02 and in the regions other than the epitopes*

Epitope[a]	Position	sN[b]	Cn[b]	sS[b]	Cs[b]	dn (Cn/sN)/ds (Cs/sS)[b]	p Value[c]
No epitope[d]		59,979.7	75.3	20,120.3	37.7	0.67	6.2×10^{-6}
Tax 11–19	11–19	9,964.7	19	4,453.3	8	1.06	0.800
XN3	21–35	18,149.6	23	5,880.4	6	1.24	0.340
XN4	31–45	16,642.3	20	7,387.7	11	0.81	0.303
XN9	80–95	18,419.9	17	7,212.1	9	0.74	0.167
XN11	101–115	18,152.4	31	5,877.6	9	1.12	0.561
XN12	111–122	14,507.0	7	4,717.0	3	0.76	0.418
All epitopes[e]		85,871.1	98	31,074.9	38	0.93	0.484

[a] CTL epitopes in amino acid position 1–133 of HTLV-I Tax have been reported by epitope mapping.

[b] The total numbers of synonymous (Cs) and nonsynonymous substitutions (Cn) independently occurring in three patients were summed in each codon site. The total numbers of Cs and Cn were counted in the T cell epitopes and the remaining regions in *tax* genes. The total numbers of synonymous (sS) and nonsynonymous sites (sN) were computed in the regions of compared sequences. The dn is a value of Cn divided by sN. The ds is a value of Cs divided by sS.

[c] Given the null hypothesis that ds equals dn, the p value was estimated by the two-tailed χ^2 test.

[d] No epitope indicates the regions other than the epitopes.

[e] All epitopes indicate all positions of epitopes above. The total numbers of synonymous (Cs) and nonsynonymous (Cn) independently occurring in three patients were summed in each codon site.

gated on forward and side scatter image. Ten thousand CD8$^+$ cells were further gated and the proportion of IFN-γ^+ cells in the CD8$^+$ cell population was analyzed. The frequency of peptide-specific CD8$^+$ T cells was obtained by subtracting the percentage of IFN-γ^+ cells without peptide from that with a peptide. The relative frequency of variant epitope-specific T cells to the frequency of Tax 11–19-specific T cells was given by the following formula: (frequency of variant epitope-specific T cells)/(frequency of Tax 11–19-specific T cells) × 100.

Results

Detection of positive selection pressures to tax *gene*

We sequenced 146 clones of the *tax* gene from patient 31, 152 clones from patient 38, and 147 clones from patient 48. The phylogenetic tree of all cloned sequences was constructed to examine the phylogenetic relationship of the *tax* gene isolated from three patients. The sequences isolated from a patient would be clustered if the sequences isolated from a patient evolve in a specific pattern. However, the sequences isolated from each patient were not clustered in the resulting phylogenetic tree, indicating that there are not patient-specific variant viruses in the three patients. We depicted the amino acid replacements in the Tax protein with the previously reported CTL epitopes in HLA-A*02 HAM/TSP patients in Fig. 1. The consensus *tax* sequence in each patient was the same to the ATK-1 sequence first reported (28), which was referred to as the prototype amino acid sequence in Fig. 1. The number of amino acid replacement at each position was 1.36 ± 1.53 (mean ± SD). Amino acid replacements over 4.42 (mean + 2 SD) were observed at amino acid positions 14, 20, 29, 43, and 54. The replacements at 14, 29, and 43 were within the defined epitopes Tax 11–19, XN3, and XN4, respectively. However, the replacements at 20 and 54 were not within the epitopes. Overall, it is unclear whether amino acid replacements significantly occur in CTL epitope regions in this analysis.

We next compared synonymous changes with nonsynonymous changes for the epitopes and the remaining (no epitope) regions in the sequenced *tax* genes. We used all sequence data for the calculation (Table I). It is generally accepted that the dn/ds ratio over 1 indicates positive selection pressure (11). In the present study, the dn/ds ratio over 1 was observed in the epitopes of Tax 11–19, XN3, and XN11. It is likely that positive selection pressure occurred in these regions, but these were not significant. In epitopes XN4, XN9, and XN12, the dn/ds ratios were <1. The ratio of all the epitope regions was higher (0.93) than that of the nonepitope regions (0.67); however, the ratio was <1. In contrast, ds was significantly higher than dn in the nonepitope regions. It indicated that the nonepitope region preferred synonymous changes to nonsynonymous changes, and was conservative in a protein level. Taken together, this suggests the possibility that positive selection pressures occur on some of the CTL epitope regions.

To further clarify whether positive selection pressure specifically occurs at the CTL epitope sites, we examined selective pressures along the *tax* genes, in which the significant test of biased synonymous and nonsynonymous substitutions was performed in a sliding window of five amino acids (Fig. 2). Positive pressures were demonstrated in four regions with statistical significance ($p < 0.05$): aa 11–15, aa 43–47, aa 56–64, and aa 105–113. Except for the aa 56–64 region, the other three regions were consistent with the epitope regions: Tax 11–19, XN4, and XN11, respectively. We found a cluster indicative of positive selection pressures at aa 56–64, however, we could not find any CTL epitopes restricted to HLA-A*02 in the literature (19–21).

No accumulation of any variant viruses of HTLV-I

Because we found that the three epitopes of Tax 11–19, XN4, and XN11 were exposed to antiviral pressures by CTL in the patients with HAM/TSP, we investigated whether some variant epitopes accumulate during the time course of the disease in these epitopes.

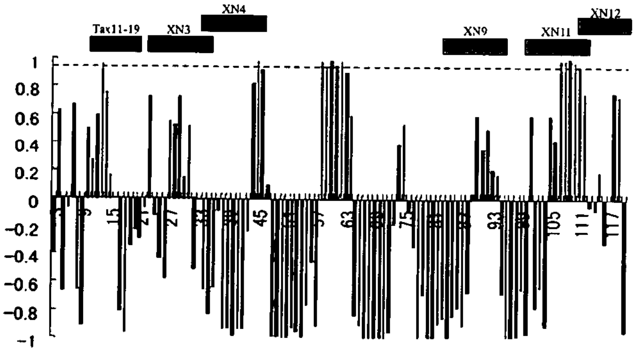


FIGURE 2. Detection of positive selection pressures on the HTLV-I *tax* gene. The figure indicates the distribution of the value (1-P) for natural selection. When dn is larger than ds, the value is indicated above the abscissa, whereas in the opposite situation, below the abscissa. Light black columns that are over the dashed line indicate positively selected sites ($p <$ 0.05 by the Fisher's exact test). The abscissa indicates the amino acid positions. This analysis was performed in a sliding window of five amino acids, and the dn or ds for the sequence is expressed at the middle position of the five amino acids. The transverse bars indicate all the known CTL epitopes which can bind to HLA-A*02.

Table II. *Longitudinal analysis of amino acid replacements in CTL epitopes, Tax 11–19, XN4, and XN11*[a]

		Tax 11–19		XN4		XN11	
Patient	Date	Amino acid sequence[c]	Frequency[b]	Amino acid sequence	Frequency	Amino acid sequence	Frequency
31	Jun/91	LLFGYPVYV	48/49	ISGGLCSARLHRHAL	46/49	IPPSFLQAMRKYSPF	48/49
		---R-----	1/49	F--------------	1/49	------R--------	1/49
				-------T-------	1/49		
				------------R--	1/49		
	Mar/97	LLFGYPVYV	45/48	ISGGLCSARLHRHAL	46/47	IPPSFLQAMRKYSPF	45/48
		P--------	1/48	----P----------	1/47	V--------------	1/48
		--Y------	1/48			----S----------	1/48
		---R-----	1/48			---------H-----	1/48
	Mar/99	LLFGYPVYV	48/49	ISGGLCSARLHRHAL	46/49	IPPSFLQAMRKYSPF	46/49
		---H-----	1/49	--E---------R--	1/49	T--------------	1/49
				-------T-------	1/49	------R--------	1/49
				------------Y--	1/49	-------T-------	1/49
38	May/96	LLFGYPVYV	48/51	ISGGLCSARLHRHAL	48/49	IPPSFLQAMRKYSPF	49/51
		--L------	1/51	----P----------	1/49	----S----------	1/51
		------I--	2/51			--------T------	1/51
	Feb/97	LLFGYPVYV	52/52	ISGGLCSARLHRHAL	48/51	IPPSFLQAMRKYSPF	47/51
				----------R----	1/51	H---X[d]	3/51
				--------------Q	2/51	--------------S	1/51
	Jul/99	LLFGYPVYV	39/49	ISGGLCSARLHRHAL	46/49	IPPSFLQAMRKYSPF	46/49
		-P-------	1/49	-----R---------	1/49	----Y----------	1/49
		--L------	2/49	------------Y--	1/49	--------L------	2/49
		---R-----	4/49	--------------P	1/49		
		----N----	1/49				
		----H----	1/49				
		-------H-	1/49				
48	May/93	LLFGYPVYV	45/48	ISGGLCSARLHRHAL	45/47	IPPSFLQAMRKYSPF	46/47
		---R--A--	1/48	T-R--------X	1/47	---------H-----	1/47
		----F----	1/48	--------------P	1/47		
		-------C-	1/48				
	Oct/96	LLFGYPVYV	48/51	ISGGLCSARLHRHAL	49/51	IPPSFLQAMRKYSPF	48/50
		---R-----	1/51	-----R---------	1/51	T--------------	1/50
		-----L---	1/51	------------Y--	1/51	--L------------	1/50
		--------E	1/51				
	Feb/00	LLFGYPVYV	47/48	ISGGLCSARLHRHAL	47/48	IPPSFLQAMRKYSPF	46/48
		P--------	1/48	-F-------------	1/48	------R--------	1/48
						--------------S	1/48

[a] At three time points in three HLA-A*0201 HAM/TSP patients, HTLV-I proviral DNA from the PBMC were subcloned and sequenced. The predicted amino acid sequences from the DNA sequences (*c*) and their frequencies (*b*) are displayed. Tax 11–19, XN4, and XN11 are CTL epitopes restricted to HLA-A*02, in which positive selection pressures were significantly detected as shown in Fig. 2. The amino acid sequence in the upper row in each sample indicates a prototype. X indicates a deletion or frame shift at the corresponding DNA sequence (*d*).

We summarized the frequencies of variant epitopes of Tax 11–19, XN4, and XN11 in Table II. In the Tax 11–19 epitope, the glycine to arginine change at position 4 was frequently observed in the three patients (1.8% in all the clone sequenced). However, no variant viruses accumulated in the time course. The prototype epitope (LLFGYPVYV) was predominant throughout the time course in all patients. In the XN4 epitope, amino acid replacements randomly occurred with little preferential replacements at positions 13 and 15, and the prototype sequence was predominant. In the XN11 epitope, replacements randomly occurred with small clusters around at positions 1 and 7, and the prototype amino acid sequence was predominant during the entire time course. Consequently, there was no accumulation of any variant viruses that escaped from the immune system in these patients.

Detection of variant epitope-specific CD8$^+$ T cells

We questioned whether no accumulation of variant viruses results from increasing CTL responses to the variant virus. Therefore, we investigated the frequency of variant virus-specific T cells during the time course of the disease. We focused on Tax 11–19 replacements, because the Tax 11–19 peptide was reported to be a strong immunodominant epitope in patients with HAM/TSP, and indeed in this study we detected the positive selection pressure in this region (7, 29). We synthesized several variant epitope peptides of

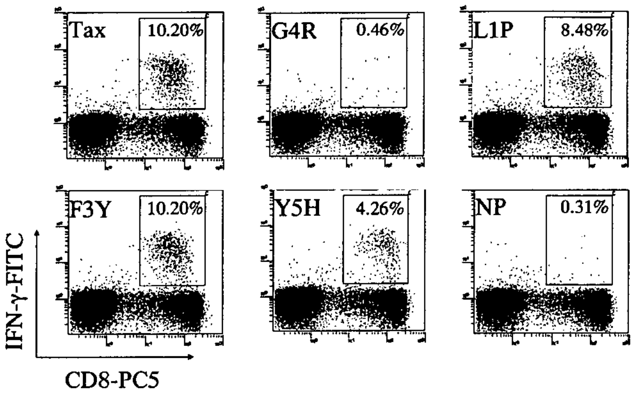


FIGURE 3. Representative detection of variant epitope-specific CD8$^+$ T cells in patient 31 in March 1999 by flow cytometry. The naturally occurring variant epitopes derived from Tax 11–19 were synthesized and used to detect variant epitope-specific CD8$^+$ T cells by intracellular IFN-γ staining. The number in the box is the Ag-specific T cell frequency in CD8$^+$ T cell population. The name of the variant epitope is designated by the amino acid replacement, i.e., G4R indicates that glycine at position 4 of the Tax 11–19 is substituted by arginine. NP indicates spontaneous IFN-γ production with APCs without any peptides.

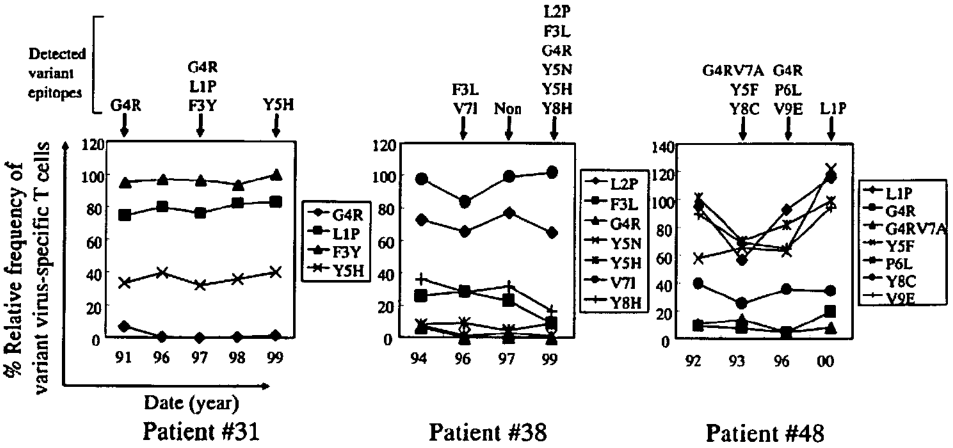


FIGURE 4. Longitudinal analysis of variant epitope-specific CD8$^+$ T cells in the patients. The *x*- and *y*-axis indicate date and relative frequency of variant epitope-specific CD8$^+$ T cells, respectively. The relative frequency was obtained by dividing the variant epitope-specific T cell frequency in the CD8$^+$ T cell population by the Tax 11–19-specific T cell frequency in the population. The variant epitopes above the boxes are indicated at the time point when the epitopes were detected in each patient.

Tax 11–19 according to the sequence data. In each patient, the variant epitopes emerged during the time course was tested to be recognized by CD8$^+$ T cells. We performed intracellular IFN-γ detection, where variant epitope-specific CD8$^+$ T cells were detected by their recognition of the epitopes loaded on APCs. We used no peptide-loaded APCs to evaluate background IFN-γ production, and influenza virus M1 peptide as a peptide control. The background IFN-γ-positive cells in CD8$^+$ cells from these patients were <0.49%, and the positive cells for M1 peptide were 1.61–5.41% in patient 31, 0.32–0.68% in patient 38, and 0.35–0.57% in patient 48 (data not shown). As shown in Fig. 3, variant epitope-specific CD8$^+$ T cells were detected; Ag-specific T cells in the CD8$^+$ T cell population were 10.20% for Tax 11–19 and for F3Y; to a lesser extent, L1P and Y5H were recognized by the T cells, however, G4R was rarely recognized. According to the formula described in *Materials and Methods*, we calculated relative frequency of variant epitope-specific T cells to Tax 11–19-specific T cells and depicted this in Fig. 4. In patient 31, F3Y peptide was recognized at the same level as Tax 11–19 as ~100%, L1P and Y5H were moderately recognized, and G4R was rarely recognized by the CD8$^+$ T cells. The relative frequencies of variant virus-specific T cells were considerably stable despite the emergence of variant viruses during the time course. In patient 38, many mutants emerged in 1999, whereas the relative frequencies of variant virus-specific T cells were almost the same. Interestingly, G4R was also rarely recognized by the T cells as in patient 31. Y5H recognition was different between these patients: ~40% in patient 31, whereas <10% in patient 38. In patient 48, the relative frequencies fluctuated; recognition of all of the variant viruses was relatively low in 1993, whereas the recognition of all of the variant viruses, except G4RV7A, was relatively high in 2000. Overall, the apparent increase of variant virus-specific T cells was not observed in association with the appearance of any variant viruses in these patients.

Discussion

We demonstrated that the positive selection pressures were detected in three of six CTL epitopes in HTLV-I Tax and the amino acid replacements were preferentially observed in these regions. There was no accumulation of any variant viruses in the time course and the proportion of variant virus-specific T cells was stable.

The positive selection pressures detected indicate that the CTL exert an effect to eliminate the virus in vivo. This is consistent with the previous report, where mutations in CTL epitope-coding regions occur significantly more frequently in HLA-A2-positive subjects than in HLA-A2-negative subjects, which suggests that HLA-A2-restricted, Tax-specific CTL induce in vivo antiviral pressures (30). The detection of positive selection pressures in Tax 11–19 is consistent with previous reports; Tax 11–19 has been demonstrated as an immunodominant epitope, which induces strong CTL responses both in HLA-A*02 HAM/TSP patients and the carriers (7, 8, 18, 29). In HIV infection, a virus with variant epitope easily escapes from the host immune system and predominates during the time course (14–16). This phenomenon has made it difficult to establish an effective CTL vaccine. However, in HTLV-I infection, although there were antiviral pressures by CTL in vivo, a virus with variant epitope did not become dominant during the time course. This may be an advantage in establishing an effective CTL vaccine. The three epitopes, where positive selection pressure and no escape variants were observed, could be candidates in designing a CTL vaccine.

The rate of amino acid replacements in HTLV-I Tax is higher than previously considered. In patient 38, the rate in Tax 11–19 ranged from 0% (in February 1997 in Table II) to 20.4% (in July 1999), suggesting that the virus actively replicates in vivo. However, any variant virus did not become dominant over the prototype virus during 3–8 years. The relative frequency of T cells specific for these peptides did not significantly change during the time course as shown in Fig. 4, and the frequencies of T cells specific for G4R (in patient 31, 38, and 48), Y5H (in patient 38), and Y5N (in patient 38) were constantly low. More importantly, the frequency of G4R-specific T cells was <10% and did not increase during the time course in all patients. These results suggest that nonaccumulation of variant viruses is not due to CTL responses. These raise the question of why the variant viruses do not become predominant, especially variants G4R, Y5H, and Y5N, which were rarely recognized by CTL (Fig. 4). Tax protein is a viral regulatory protein, which facilitates viral replication via trans-activator function that up-regulates the numerous promoter genes including its long terminal repeat promoter, IL-2R α promoter (31). Niewiesk et al. (30) explored whether naturally occurring variants of HTLV-I Tax impair the transactivation function, and they demonstrated that most of the amino acid substitutions in Tax protein severely reduced its ability to transactivate three promoters: the HTLV-I long terminal repeat, the c-*fos* promoter, and the IL-2R α-chain promoter (30). The replacement of the *tax* gene that codes for the G4R epitope is frequently observed both in their study and ours. In their study, the G4R replacement strongly reduces trans-activator function (30). The reduction of transactivator activity by a replacement is reported not only in the Tax 11–19 epitope but also in XN4 and XN11 (30). Furthermore, artificial random mutagenesis of the *tax* gene, which introduces amino acid substitutions in Tax protein, abolishes the Tax regulatory functions (32). Therefore, some replacements in the regulatory protein Tax may impair the transactivator function for viral replication, which may lead to nonpredominance of the variant viruses, even if they are rarely recognized by CTL. Recently, reversion of CTL escape-variant SIV to the original virus is reported in newly infected animals in the absence of selective pressure by CTL, which suggests that viral evolution is a result of the equilibrium between viral fitness for replication and viral escape from the immune system (33, 34). Our data may support this hypothesis.

It is reported that the replacement rate is lower in HTLV-I than in HIV in vivo (35). Retroviruses can replicate by two ways within infected individuals: mitotic division of virus-infected cells and infectious spread among cells via reverse transcriptase. Although HIV spread mainly by free viral infection within infected individuals, it has been considered that HTLV-I mainly spread by mitotic replication, because infectivity of HTLV-I is extremely low in vitro and there is no evidence that HTLV-I is transmissible from HTLV-I-infected individuals to another person by blood component-depleted lymphocytes. Furthermore, inverse PCR analysis, which can distinguish each HTLV-I-infected T cell clone, reveals that clonal expansion of several infected cells is a common feature of HTLV-I infection (36). However, in our study, the proportion of mutant virus reached up to 20.4% (Table II, patient 38 on Jul/99). Moreover, viral sequence analysis revealed positive selection pressures, which is generally detected when retroviruses replicate by reverse transcriptase and not by mitosis of the infected cells. Thus, although HTLV-I may mainly spread via mitotic replication, replication via reverse transcriptase can play a role in increasing proviral load at least in some individuals. Quantification of the ratio of mitotic replication vs replication via reverse transcriptase may be important in using reverse transcriptase inhibitors for therapeutic purposes.

We found positive selection pressure around position 54 in the Tax protein (Fig. 2); however, HLA-A*02-restricted CTL epitopes have not been reported in these regions. Although the reason why the pressure was observed in the region is unclear, there is a possibility that positive selection pressure by CTL occurs on these amino acids; another T cell epitope restricted to an HLA allele other than HLA-A*02 may be found in these patients.

In conclusion, Tax-specific CTL act as killer cells, which induce positive selection pressure to HTLV-I in vivo. The naturally occurring variant viruses do not become predominant in the viral population unlike that seen in HIV infection. Therefore, these epitopes may be candidate targets for HTLV-I vaccine development. A search for a viral protein, which includes CTL epitopes and is essential for viral replication, may be important in designing a CTL vaccine for chronic viral infections.

Acknowledgments

We thank Drs. Arlene R. Ng and Moe Moe Aye for the critical reading of the manuscript.

Disclosures

The authors have no financial conflict of interest.

References

1. Uchiyama, T., J. Yodoi, K. Sagawa, K. Takatsuki, and H. Uchino. 1977. Adult T-cell leukemia: clinical and hematologic features of 16 cases. *Blood* 50: 481–492.
2. Osame, M., K. Usuku, S. Izumo, N. Ijichi, H. Amitani, A. Igata, M. Matsumoto, and M. Tara. 1986. HTLV-I associated myelopathy, a new clinical entity. *Lancet* 1: 1031–1032.
3. Umehara, F., S. Izumo, M. Nakagawa, A. T. Ronquillo, K. Takahashi, K. Matsumuro, E. Sato, and M. Osame. 1993. Immunocytochemical analysis of the cellular infiltrate in the spinal cord lesions in HTLV-I-associated myelopathy. *J. Neuropathol. Exp. Neurol.* 52: 424–430.
4. Osame, M., M. Matsumoto, K. Usuku, S. Izumo, N. Ijichi, H. Amitani, M. Tara, and A. Igata. 1987. Chronic progressive myelopathy associated with elevated antibodies to human T-lymphotropic virus type I and adult T-cell leukemialike cells. *Ann. Neurol.* 21: 117–122.
5. Nagai, M., K. Usuku, W. Matsumoto, D. Kodama, N. Takenouchi, T. Moritoyo, S. Hashiguchi, M. Ichinose, C. R. Bangham, S. Izumo, and M. Osame. 1998. Analysis of HTLV-I proviral load in 202 HAM/TSP patients and 243 asymptomatic HTLV-I carriers: high proviral load strongly predisposes to HAM/TSP. *J. Neurovirol.* 4: 586–593.
6. Takenouchi, N., Y. Yamano, K. Usuku, M. Osame, and S. Izumo. 2003. Usefulness of proviral load measurement for monitoring of disease activity in individual patients with human T-lymphotropic virus type I-associated myelopathy/ tropical spastic paraparesis. *J. Neurovirol.* 9: 29–35.
7. Jacobson, S., H. Shida, D. E. McFarlin, A. S. Fauci, and S. Koenig. 1990. Circulating $CD8^+$ cytotoxic T lymphocytes specific for HTLV-I pX in patients with HTLV-I associated neurological disease. *Nature* 348: 245–248.
8. Kannagi, M., S. Harada, I. Maruyama, H. Inoko, H. Igarashi, G. Kuwashima, S. Sato, M. Morita, M. Kidokoro, M. Sugimoto, et al. 1991. Predominant recognition of human T cell leukemia virus type I (HTLV-I) pX gene products by human $CD8^+$ cytotoxic T cells directed against HTLV-I-infected cells. *Int. Immunol.* 3: 761–767.
9. Parker, C. E., S. Daenke, S. Nightingale, and C. R. Bangham. 1992. Activated, HTLV-1-specific cytotoxic T-lymphocytes are found in healthy seropositives as well as in patients with tropical spastic paraparesis. *Virology* 188: 628–636.
10. Tomaru, U., Y. Yamano, M. Nagai, D. Maric, P. T. Kaumaya, W. Biddison, and S. Jacobson. 2003. Detection of virus-specific T cells and $CD8^+$ T-cell epitopes by acquisition of peptide-HLA-GFP complexes: analysis of T-cell phenotype and function in chronic viral infections. *Nat. Med.* 9: 469–476.
11. Nei, M., and T. Gojobori. 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol. Biol. Evol.* 3: 418–426.
12. Niewiesk, S., S. Daenke, C. E. Parker, G. Taylor, J. Weber, S. Nightingale, and C. R. Bangham. 1994. The transactivator gene of human T-cell leukemia virus type I is more variable within and between healthy carriers than patients with tropical spastic paraparesis. *J. Virol.* 68: 6778–6781.
13. Zinkernagel, R. M., and P. C. Doherty. 1974. Restriction of in vitro T cell-mediated cytotoxicity in lymphocytic choriomeningitis within a syngeneic or semiallogeneic system. *Nature* 248: 701–702.
14. Phillips, R. E., S. Rowland-Jones, D. F. Nixon, F. M. Gotch, J. P. Edwards, A. O. Ogunlesi, J. G. Elvin, J. A. Rothbard, C. R. Bangham, C. R. Rizza, et al. 1991. Human immunodeficiency virus genetic variation that can escape cytotoxic T cell recognition. *Nature* 354: 453–459.
15. Borrow, P., H. Lewicki, X. Wei, M. S. Horwitz, N. Peffer, H. Meyers, J. A. Nelson, J. E. Gairin, B. H. Hahn, M. B. Oldstone, and G. M. Shaw. 1997. Antiviral pressure exerted by HIV-1-specific cytotoxic T lymphocytes (CTLs) during primary infection demonstrated by rapid selection of CTL escape virus. *Nat. Med.* 3: 205–211.
16. Evans, D. T., D. H. O'Connor, P. Jing, J. L. Dzuris, J. Sidney, J. da Silva, T. M. Allen, H. Horton, J. E. Venham, R. A. Rudersdorf, et al. 1999. Virus-specific cytotoxic T-lymphocyte responses select for amino-acid variation in simian immunodeficiency virus Env and Nef. *Nat. Med.* 5: 1270–1276.
17. Kubota, R., Y. Furukawa, S. Izumo, K. Usuku, and M. Osame. 2003. Degenerate specificity of HTLV-1-specific $CD8^+$ T cells during viral replication in patients with HTLV-1-associated myelopathy (HAM/TSP). *Blood* 101: 3074–3081.
18. Parker, C. E., S. Nightingale, G. P. Taylor, J. Weber, and C. R. Bangham. 1994. Circulating anti-Tax cytotoxic T lymphocytes from human T-cell leukemia virus type I-infected people, with and without tropical spastic paraparesis, recognize multiple epitopes simultaneously. *J. Virol.* 68: 2860–2868.
19. Daenke, S., A. G. Kermode, S. E. Hall, G. Taylor, J. Weber, S. Nightingale, and C. R. Bangham. 1996. High activated and memory cytotoxic T-cell responses to HTLV-1 in healthy carriers and patients with tropical spastic paraparesis. *Virology* 217: 139–146.
20. Koenig, S., R. M. Woods, Y. A. Brewah, A. J. Newell, G. M. Jones, E. Boone, J. W. Adelsberger, M. W. Baseler, S. M. Robinson, and S. Jacobson. 1993. Characterization of MHC class I restricted cytotoxic T cell responses to tax in HTLV-1 infected patients with neurologic disease. *J. Immunol.* 151: 3874–3883.
21. Pique, C., F. Connan, J. P. Levilain, J. Choppin, and M. C. Dokhelar. 1996. Among all human T-cell leukemia virus type 1 proteins, tax, polymerase, and envelope proteins are predicted as preferential targets for the HLA-A2-restricted cytotoxic T-cell response. *J. Virol.* 70: 4919–4926.
22. Felsenstein, J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol.* 17: 368–376.
23. Hartigan, J. A. 1973. Minimum evolution fits to a given tree. *Biometrics* 29: 53–65.
24. Zhang, J., S. Kumar, and M. Nei. 1997. Small-sample tests of episodic adaptive evolution: a case study of primate lysozymes. *Mol. Biol. Evol.* 14: 1335–1338.
25. Suzuki, Y., and T. Gojobori. 1999. A method for detecting positive selection at single amino acid sites. *Mol. Biol. Evol.* 16: 1315–1328.

26. Suzuki, Y., T. Gojobori, and M. Nei. 2001. ADAPTSITE: detecting natural selection at single amino acid sites. *Bioinformatics* 17: 660–661.

27. Utz, U., S. Koenig, J. E. Coligan, and W. E. Biddison. 1992. Presentation of three different viral peptides, HTLV-1 Tax, HCMV gB, and influenza virus M1, is determined by common structural features of the HLA-A2.1 molecule. *J. Immunol.* 149: 214–221.

28. Seiki, M., S. Hattori, Y. Hirayama, and M. Yoshida. 1983. Human adult T-cell leukemia virus: complete nucleotide sequence of the provirus genome integrated in leukemia cell DNA. *Proc. Natl. Acad. Sci. USA* 80: 3618–3622.

29. Kannagi, M., H. Shida, H. Igarashi, K. Kuruma, H. Murai, Y. Aono, I. Maruyama, M. Osame, T. Hattori, H. Inoko, et al. 1992. Target epitope in the Tax protein of human T-cell leukemia virus type I recognized by class I major histocompatibility complex-restricted cytotoxic T cells. *J. Virol.* 66: 2928–2933.

30. Niewiesk, S., S. Daenke, C. E. Parker, G. Taylor, J. Weber, S. Nightingale, and C. R. Bangham. 1995. Naturally occurring variants of human T-cell leukemia virus type I Tax protein impair its recognition by cytotoxic T lymphocytes and the transactivation function of Tax. *J. Virol.* 69: 2649–2653.

31. Yoshida, M., J. Inoue, J. Fujisawa, and M. Seiki. 1989. Molecular mechanisms of regulation of HTLV-1 gene expression and its association with leukemogenesis. *Genome* 31: 662–667.

32. Smith, M. R., and W. C. Greene. 1990. Identification of HTLV-1 tax transactivator mutants exhibiting novel transcriptional phenotypes. *Genes Dev.* 4: 1875–1885.

33. Friedrich, T. C., E. J. Dodds, L. J. Yant, L. Vojnov, R. Rudersdorf, C. Cullen, D. T. Evans, R. C. Desrosiers, B. R. Mothe, J. Sidney, et al. 2004. Reversion of CTL escape-variant immunodeficiency viruses in vivo. *Nat. Med.* 10: 275–281.

34. Leslie, A. J., K. J. Pfafferott, P. Chetty, R. Draenert, M. M. Addo, M. Feeney, Y. Tang, E. C. Holmes, T. Allen, J. G. Prado, et al. 2004. HIV evolution: CTL escape mutation and reversion after transmission. *Nat. Med.* 10: 282–289.

35. Hanada, K., Y. Suzuki, and T. Gojobori. 2004. A large variation in the rates of synonymous substitution for RNA viruses and its relationship to a diversity of viral infection and transmission modes. *Mol. Biol. Evol.* 21: 1074–1080.

36. Wattel, E., J. P. Vartanian, C. Pannetier, and S. Wain-Hobson. 1995. Clonal expansion of human T-cell leukemia virus type I-infected cells in asymptomatic and symptomatic carriers without malignancy. *J. Virol.* 69: 2863–2868.

Compensatory Change of Interacting Amino Acids in the Coevolution of Transcriptional Coactivator MBF1 and TATA-Box–Binding Protein

Qing-Xin Liu,[1] Naomi Nakashima-Kamimura,†[1] Kazuho Ikeo,* Susumu Hirose,† and Takashi Gojobori**

*Center for Information Biology and DNA Data Bank of Japan, National Institute of Genetics, Mishima, Shizuoka, Japan; and †Department of Developmental Genetics, National Institute of Genetics, Mishima, Shizuoka, Japan

To elucidate the transcriptional regulation in eukaryotic genome network, it is important to understand coevolution of transcription factors, transcriptional coactivators, and TATA-box–binding protein (TBP). In this study, coevolution of transcriptional coactivator multiprotein-bridging factor 1 and its interacting target TBP was first evaluated experimentally by examining if compensatory amino acid changes took place at interacting sites of both proteins. The experiments were conducted by identifying interaction sites and comparing the amino acids at these sites among different organisms. Here, we provide evidence for compensatory changes of transcription coactivator and its interacting target, presenting the 1st report that transcription coactivator may have undergone coevolution with TBP.

Introduction

Transcription factors are components of the regulatory network and are involved in multiple interactions with other proteins. Thus, the evolution of transcription factor families takes place within a framework defined by these interactions. Because the conserved domains within transcription factors often contain sites that mediate these interactions, their conservation most likely reflects the conservation of classes of interactions that were established early in evolution and under the limitations of tolerable (i.e., functional) changes. Many transcription factors have been evolutionarily conserved (Ge et al. 2002; Stevens et al. 2002; Taatjes et al. 2004; Bustamante et al. 2005); however, the evolutionary mechanism of transcription factors remains unclear. Multiprotein-bridging factor 1 (MBF1) is a transcriptional coactivator that mediates transcriptional activation by bridging a sequence-specific activator and TATA-box–binding protein (TBP) (Li et al. 1994; Takemaru et al. 1997, 1998; Liu et al. 2003; Jindra et al. 2004). Interaction between MBF1 and TBP is conserved from *Archaea* to humans (Aravind and Koonin 1999; Kabe et al. 1999; Millership et al. 2004). To understand the evolutionary mechanism of transcription coactivator, we have analyzed the evolution of MBF1 and TBP. Here, we examine whether coevolution can be evaluated by an experimental method in which we first identified interacting amino acids between 2 proteins and then carried out evolutionary analysis.

Materials and Methods

Polymerase Chain Reaction Mutagenesis

The *tbp* (*TBP* gene–containing mutations) library was made by error-prone polymerase chain reaction (PCR) as described (Lin-Goerke et al. 1997). In brief, the yeast *TBP* (y*TBP*) gene was amplified by PCR using Taq DNA polymerase in a reaction mixture containing 0.25 mM $MnCl_2$. The PCR products were then purified and inserted into a YCplac22 vector, *TRP1*-marked plasmid.

Screening for *tbp* Genes

The *Saccharomyces cerevisiae* yeast strain used for screening was N107-1 (*MATa ade2-1 ura3-1 trp1-1 leu2-3, 112 can1-100 Δtbp::LEU2* [Ycplac33-y*TBP*]). This strain has chromosomal *TBP* gene deletion replaced by an *URA3*-marked plasmid carrying the *TBP* gene. The *tbp* library was transformed into N107-1 and spread onto plates with synthetic complete (SC) media but did not contain tryptophan. Strains expressing *tbp* were grown on 5-fluoroorotic acid (5-FOA) to remove plasmid-carrying *TBP* and then shifted to submaster plates either containing aminotriazole (AT) or not containing AT. We screened strains that were AT sensitive but showed normal growth.

Glutathione *S*-transferase Pull-Down Assay

Q68L and Q68I *tbp* genes were subcloned into 6HisT-pET11d to produce TBP proteins bearing 6 histidine residues in *Escherichia coli*. His-tagged recombinant proteins were purified using a Ni-column (Novagen, San Diego, CA) and used for assay. The Glutathione *S*-transferase (GST) pull-down assay was performed using GST–MBF1 as described (Takemaru et al. 1998). Bound proteins were detected on western blots using an anti-TBP antibody.

Protein Sequence Analysis

All available sequences were obtained using the Entrez Protein database at National Center for Biotechnology Information (NCBI) (http://www.ncbi.nlm.nih.gov/). Accession numbers and species were compiled in a supplementary table S1 (Supplementary Material online). Protein-coding sequences were aligned using the ClustalX program (Thompson et al. 1997). All amino acid positions with gaps were excluded from this analysis. For phylogenetic reconstruction of TBP (Supplementary fig. 1, Supplementary Material online), the Neighbor-Joining method was used (Saitou and Nei 1987) with observed differences as implemented Njplot (Perriere and Gouy 1996). Bootstrap analysis with 1,000 replicates was used to assess the support for

[1] Equal contribution to this work.

Key words: compensatory change, coevolution, transcriptional coactivator, MBF1, TBP.

E-mail: tgojobor@genes.nig.ac.jp.

Mol. Biol. Evol. 24(7):1458–1463. 2007
doi:10.1093/molbev/msm073
Advance Access publication April 17, 2007

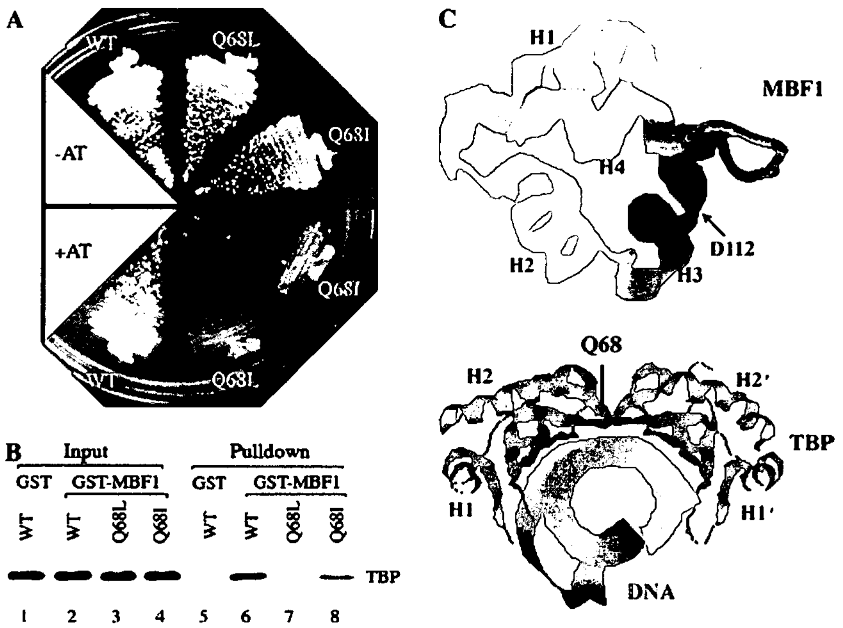


FIG. 1.—Analysis of yTBP mutants. (*A*) Q68 mutations show AT sensitivity. The *Δtbp* strain, containing the wild-type (WT) *TBP*, Q68L *tbp*, or Q68I *tbp* gene, was streaked on plates in the presence or absence of 20 mM AT and incubated for 3 days at 30 °C. (*B*) Q68 is required for binding with MBF1. Bacterially expressed and purified WT yTBP (lanes 5 and 6), Q68L (lane 7), or Q68I (lane 8) was incubated with either GST (lane 5) or GST–MBF1 (lanes 6–8). The bound yTBP was electrophoresed by Sodium dodecyl sulfate–polyacrylamide gel electrophoresis and detected with an anti-TBP antibody. Lanes 1–4 were 1/10 of the input TBP or its mutants. (*C*) The structure of *Bombyx mori* MBF1 (residues 67–146) corresponds to 73–151 of yMBF1 (top) and the structure of yTBP (residues 61–240) and DNA (bottom). D112 of yMBF1 and Q68 of yTBP are indicated by arrows. H indicates the helix motif.

tree nodes (Felsenstein 1985). Phylogenetic distribution of interaction amino acids between MBF1 and TBP is based on phylogenetic analysis of full-length TBP sequences and recent studies (Hedges 2002).

Compensatory Change Analysis

The *S. cerevisiae* yeast strain used for compensatory change of interaction amino acid analysis was N111-4A (*ade2-1 ura3-1 trp-1 leu2-3, 112 can1-100 Δmbf1::LEU2 Δtbp::LEU2* [yCplac33-TBP]). Ycplac22-TBP and Ycplac22-tbp mutant plasmids were transformed into N111-4A and spread onto plates with SC media but do not contain tryptophan. Strains expressing *tbp* were grown on 5-FOA to remove plasmid-carrying TBP. Ycplac33-MBF1 or Ycplac33-mbf1 mutant plasmids were then transformed into strains and spread onto plates with SC media but not contain Uracil. These strains were then shifted to submaster plates either containing AT or not containing AT. Point mutations of MBF1 and TBP were introduced by site-directed PCR mutagenesis.

Results

Q68 of yTBP Is Required for yMBF1 Binding

MBF1 and TBP sequences are evolutionarily conserved from *Archaea* to humans (supplementary table S1, Supplementary Material online). To obtain evidence for the coevolution of MBF1 and TBP, we identified amino acids involved in the interaction between MBF1 and TBP in the yeast *S. cerevisiae*. It has been shown that D112 of yMBF1 is necessary for the yTBP binding (Takemaru et al. 1998) although the binding site in yTBP is not known. To identify the interaction site in yTBP, we constructed a yTBP mutant library and screened mutants that are defective in the yMBF1 binding. The yMBF1 mediated GCN4-dependent transcriptional activation that is essential for derepression of the amino acid biosynthesis genes (Takemaru et al. 1998). The disruptant of yMBF1 was sensitive to AT, an inhibitor of the *HIS3* gene product. This sensitivity was also observed in the D112 yMBF1 mutant, indicating that interaction between yMBF1 and yTBP is required for the activation of *HIS3* gene transcription. We therefore screened AT-sensitive TBP mutants and obtained 4 candidates, Q68L, Q68I, R79W, and T215S. Because TBP is a general transcription factor, mutations in a site involved in a general function (e.g., DNA binding site) reduce transcription of many genes, including *HIS3* and showing AT sensitivity. To eliminate such mutations, we compared strain growth in AT-containing medium (cell requires general control non-depressible 4[GCN4] activity) and galactose-containing medium (cell requires galactose 4 activity) and found that Q68 was a specific site for the GCN4-dependent transcriptional activation (data not shown). Q68L and Q68I mutants were viable in the absence of AT (fig. 1A) and able to grow on glucose, galactose, sucrose, or inositol-free media (Supplementary fig. 1, Supplementary Material online), indicating that these mutants can achieve most TBP functions and have not destructed the TBP structure. To confirm the interaction via Q68, we performed a GST pull-down assay using a series of bacterially expressed yTBP and

GST–yMBF1 fusion proteins. GST pull-down assays with these purified proteins showed that wild-type yTBP and yMBF1 bind directly, but TBP harboring Q68I or Q68L mutation showed a significantly reduced capability of the binding to yMBF1 (fig. 1*B*). These results demonstrate that the amino acid Q68 is important for the yMBF1 binding. Whereas D112 of yMBF1 is present in the 3rd helix of the C-terminal domain, Q68 of yTBP is on top of the saddle-shaped molecule (fig. 1*C*).

Alignment Analysis of MBF1 and TBP

To understand how interacting amino acids evolve, we next analyzed the sequences of MBF1 and TBP of various organisms (figs. 2 and 3). Archaeal MBF1 contains a Zn-ribbon motif that is absent in their eukaryotic counterparts. Eukaryotic MBF1 consists of 2 structural domains; a well-structured C-terminal half that binds to TBP and a flexible N-terminal half that participates in binding to various activators (Ozaki et al. 1999). Archaeal MBF1 harbors its own DNA-binding domain (Zn-ribbon motif, fig. 2) and hence serves for a single activator. In contrast, eukaryotic MBF1 does not directly bind to DNA but interacts with various activators. Eukaryotic MBF1 seems to lose a DNA-binding motif to accommodate a variety of activator partners. In *Archaea*, the amino acid of MBF1 corresponding to yMBF1 D112 is lysine, arginine, serine, or asparagine, but it changes to aspartic acid, glutamine, or glutamic acid in eukaryotes (fig. 2). In *Archaea*, the amino acid of TBP corresponding to yTBP Q68 is glutamic acid or glutamine, whereas it changes to histidine or glutamine in eukaryotes (fig. 3). Amino acid substitution in MBF1, therefore, appears to accompany the compensatory change in TBP to maintain MBF1–TBP interaction. These results strongly suggest the coevolution of MBF1 with TBP.

Compensatory Change Analysis In Vivo

To confirm the compensatory change of interacting amino acids, we did an in vivo analysis of interacting amino acids of MBF1 and TBP in the yeast *S. cerevisiae*. We made mutants of MBF1 and TBP according to the results of the evolutionary analysis (fig. 4; Supplementary fig. 2, Supplementary Material online). The mutants TBP-68Q, MBF1-112K and TBP-68E, MBF1-112D were sensitive to 3AT (fig. 5), indicating that the interactions between MBF1 and TBP were disrupted in these mutants. As expected from the evolutionary analysis (fig. 4), the mutants TBP-68Q, MBF1-112R; TBP-68E, MBF1-112R; TBP-68E, MBF1-112N; TBP-68H, MBF1-112E; and TBP-68E, MBF1-112K were resistant to AT (fig. 5). These results suggest that the compensatory change occurred and was selected from the neutral mutations during the evolution of MBF1 and TBP.

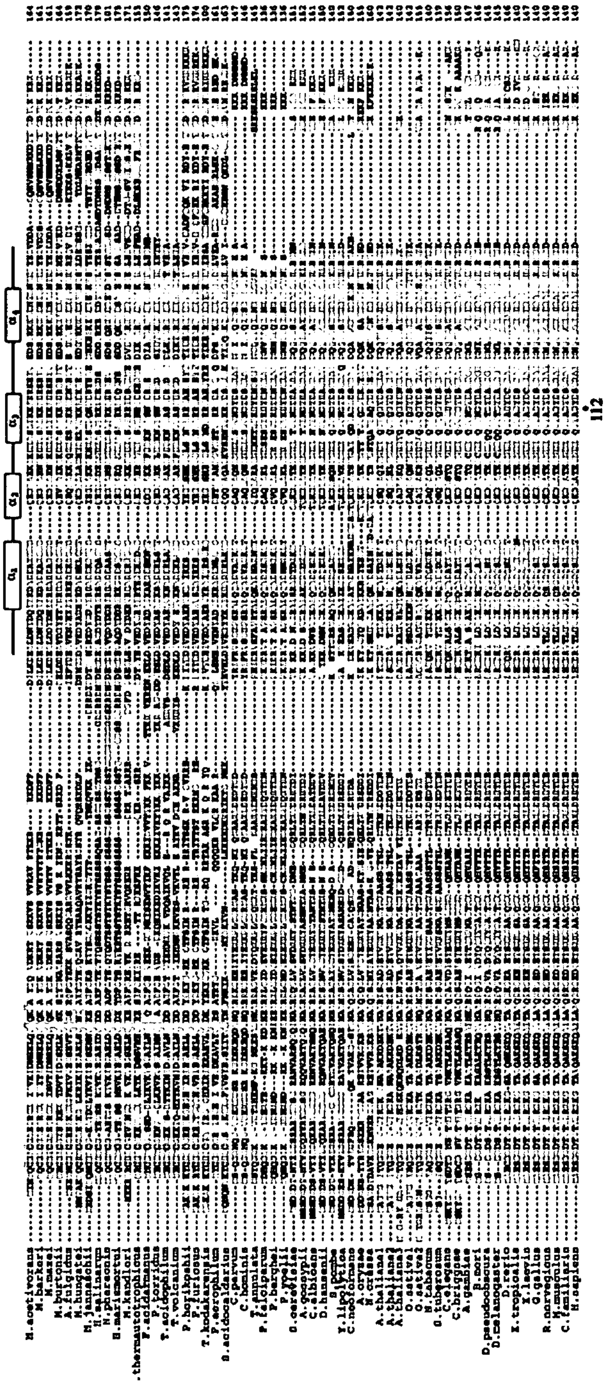


FIG. 2.—Sequence alignment of MBF1 proteins. The Zn-ribbon motif is shown by red column. Bar represents a well-structured domain, α1–α4 denotes 4 amphipathic helices. The dot indicates the amino acid bound to TBP. The number below the dot shows the amino acid position of yMBF1.

Discussion

Coevolution of MBF1 and TBP

Coevolution is a process in which an inheritable change in one entity exerts selective pressure for a change in another entity. The coevolution of proteins has been well studied (Pazos et al. 1997; Goh et al. 2000; Goh and Cohen 2002; Ramani and Marcotte 2003). If the conformation of one protein is interrupted by a mutation, a compensatory change may be selected in its interacting partner. When such

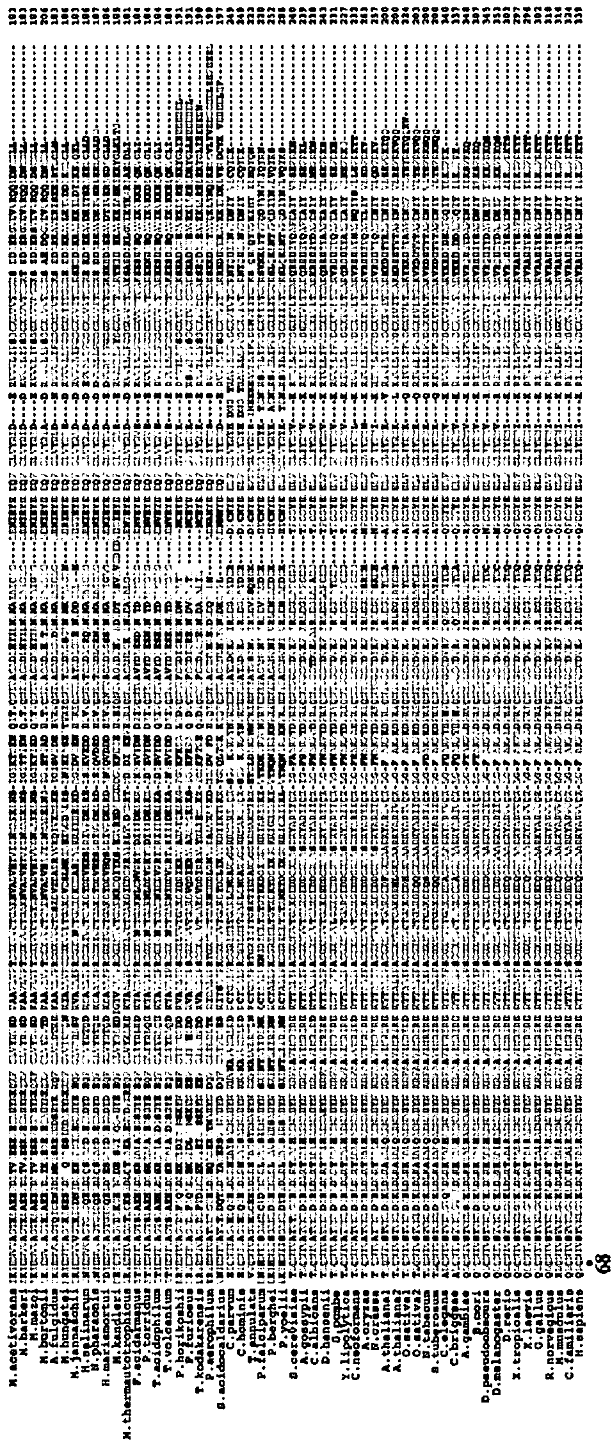
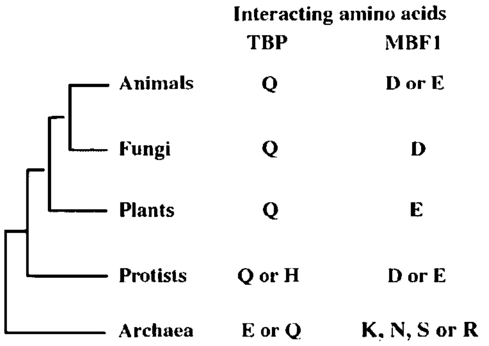


FIG. 3.—Sequence alignment of TBP proteins. Only the conserved C-terminal domain is shown. The dot indicates the amino acid bound to MBF1. The number below the dot shows the amino acid position of yTBP.

compensatory changes occur, this provides evidence of coevolution. In this study, we examined coevolution of transcription factor and coactivator for the first time and found that MBF1 coevolves with TBP. For *Archaea*, MBF1 binds to TBP through lysine, arginine, serine, or asparagine to glutamic acid interaction, or arginine–glutamine interaction. For protists, MBF1 binds to TBP through aspartic acid or glutamic acid to glutamine interaction, or glutamic acid–histidine interaction. For fungi, MBF1 binds to TBP through aspartic acid or glutamine to glutamine interaction. For plants, MBF1 binds to TBP through glutamic acid to glutamine interaction. For animals, MBF1 binds to TBP through glutamic acid or aspartic acid to glutamine interaction. As lysine does not interact with glutamine and aspartic acid does not interact with glutamic acid (fig. 5), our data indicates that an amino acid substitution in one protein results in giving selection pressure for a reciprocal change in the interacting partner. These findings suggest that a compensatory change of interacting amino acids were selected during the coevolution of MBF1 and TBP.



	Interacting amino acids	
	TBP	MBF1
Animals	Q	D or E
Fungi	Q	D
Plants	Q	E
Protists	Q or H	D or E
Archaea	E or Q	K, N, S or R

FIG. 4.—Phylogenetic distribution of interacting amino acids of MBF1 and TBP. The cladogram of relationship is based on recent studies (Hedges 2002).

Why Is Interaction between MBF1 and TBP Conserved?

MBF1 is conserved among all organisms in which TBP is used as the general transcription factor. The coactivator is preserved even in a parasitic protozoan *Cryptosporidium parvum* where many essential genes are lost from its genome and their functions are supplied by the host counterparts (Abrahamsen et al. 2004). This study demonstrated the coevolution of MBF1 and the essential protein TBP. All these findings suggest the importance of MBF1. Nevertheless, null mutants of MBF1 are viable in both yeast and *Drosophila* under laboratory conditions. Does this contradict the neutral theory of evolution (Kimura 1955), which predicts the importance of conserved genes? The answer is no. Recently, studies revealed diverse biological function of MBF1. Yeast MBF1 supports the GCN4-dependent activation of the *HIS3* gene (Takemaru et al. 1998), and *Drosophila* MBF1 serves as a coactivator of basic leucine zipper protein Tracheae defective during morphogenesis of the tracheal and nervous systems (Liu et al. 2003). *Drosophila* MBF1 also interacts with AP-1 to preserve redox-dependent AP-1 activity during oxidative stress (Jindra et al. 2004). Rat MBF1 has been isolated as a calmodulin-associated peptide 19 (Smith et al. 1998), and human MBF1 has been identified as endothelial differentiation–related factor 1 (Mariotti et al. 2000). Tomato MBF1 is induced immediately and transiently in ethylene-treated late immature fruit (Zegzouti et al. 1999). Potato MBF1 is upregulated

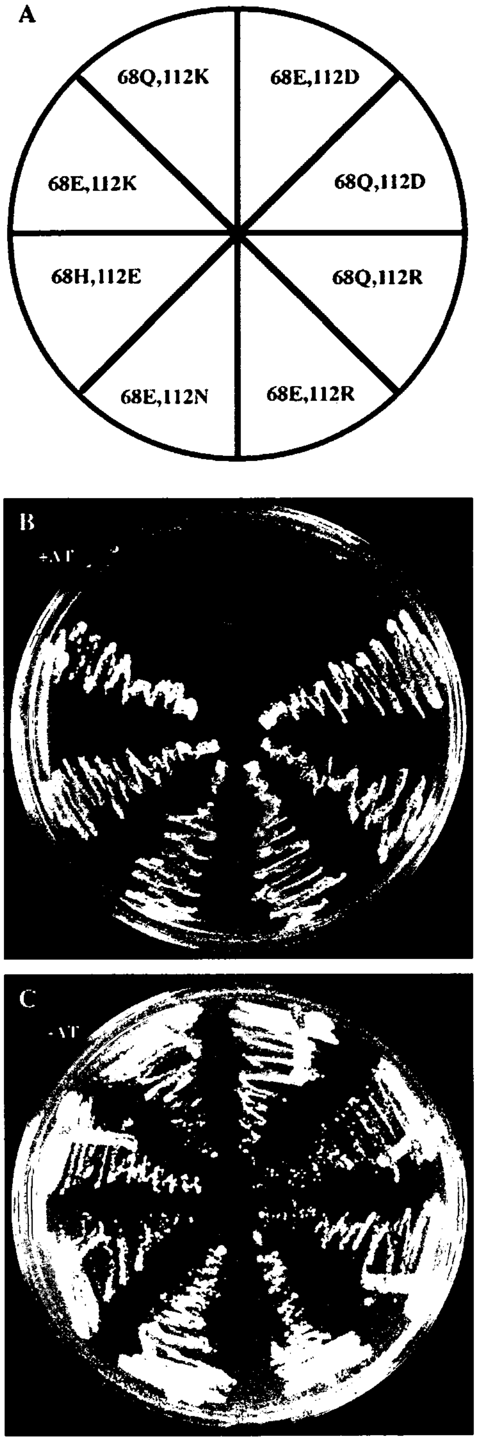


FIG. 5.—Functional analysis of mutants in yeast. (*A*) Schematic illustration of mutants using this study, yeast wild type (68Q, 112D) as a control. (*B* and *C*) Growth of yeast strains in a synthetic medium without histidine in the presence (*B*) or in the absence (*C*) of the inhibitor 20 mM 3AT.

during fungal attack, upon wounding, and by treatment with salicylic acid and the ethylene precursor ethephon (Godoy et al. 2001). Tobacco MBF1 is induced by the combined effect of drought stress and heat shock (Rizhsky et al. 2002). Therefore, the interaction between MBF1 and TBP appears to be essential in the real world where organisms are subject to nutrient starvation and various kinds of stresses, and proper differentiation timing is critical for life.

Supplementary Material

Supplementary table S1 and Supplementary figures 1 and 2 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

We thank Dr Naruya Saitou and Dr Yoshiyuki Suzuki for helpful discussions. This work was supported by Grants-in-aid for Scientific Research from the Ministry of Education, Science, Sports, Culture and Technology of Japan.

Literature Cited

Abrahamsen MS, Templeton TJ, Enomoto S, et al. (20 couthors). 2004. Complete genome sequence of the Apicomplexan, *Cryptosporidium parvum*. Science. 304:441–445.

Aravind L, Koonin EV. 1999. DNA-binding proteins and evolution of transcription regulation in the *archaea*. Nucleic Acids Res. 27:4658–4670.

Bustamante CD, Fledel-Alon A, Williamson S, et al. (14 couthors). 2005. Natural selection on protein-coding genes in the human genome. Nature. 437:1153–1157.

Felsenstein J. 1985. Confidence limits on phylogenies: an approach using the bootstrap. Evolution. 39:738–791.

Ge K, Guermah M, Yuan CX, Ito M, Wallberg AE, Spiegelman BM, Roeder RG. 2002. Transcription coactivator TRAP220 is required for PPAR2-stimulated adipogenesis. Nature. 417:563–567.

Godoy AV, Zanetti ME, San SB, Casalongué CA. 2001. Identification of a putative *Solanum tuberosum* transcriptional coactivator up-regulated in potato tubers by *Fusarium solani* f. sp. *eumartii* infection and wounding. Physiol Plant. 112: 217–222.

Goh CS, Bogan AA, Joachimiak M, Walther D, Cohen FE. 2000. Co-evolution of proteins with their interaction partners. J Mol Biol. 299:283–293.

Goh CS, Cohen FE. 2002. Co-evolutionary analysis reveals insights into protein–protein interactions. J Mol Biol. 324:177–192.

Hedges SB. 2002. The origin and evolution of model organisms. Nat Rev Genet. 3:838–849.

Jindra M, Gaziova I, Uhlirova M, Okabe M, Hiromi Y, Hirose S. 2004. Coactivator MBF1 preserves the redox-dependent AP-1 activity during oxidative stress in Drosophila. EMBO J. 23:3538–3547.

Kabe Y, Goto M, Shima D, Imai T, Wada T, Morohashi K, Shirakawa M, Hirose S, Handa H. 1999. The role of human MBF1 as a transcriptional coactivator. J Biol Chem. 274: 34196–34202.

Kimura M. 1955. Solution of a process of random genetic drift with a continuous model. Proc Natl Acad Sci USA. 41: 144–150.

Li FQ, Ueda H, Hirose S. 1994. Mediators of activation of fushi tarazu gene transcription by BmFTZ-F1. Mol Cell Biol. 14:3013–3021.

Lin-Goerke JL, Robbins DJ, Burczak JD. 1997. PCR-based random mutagenesis using manganese and reduced dNTP concentration. Biotechniques. 23:409–412.

Liu QX, Jindra M, Ueda H, Hiromi Y, Hirose S. 2003. *Drosophila* MBF1 is a co-activator for Tracheae Defective and contributes to the formation of tracheal and nervous systems. Development. 130:719–728.

Mariotti M, De Benedictis L, Avon E, Maier JAM. 2000. Interaction between endothelial differentiation-related factor-1 and calmodulin *in vitro* and *in vivo*. J Biol Chem. 275: 24047–24051.

Millership JJ, Waghela P, Cai X, Cockerham A, Zhu G. 2004. Differential expression and interaction of transcription co-activator MBF1 with TATA-binding protein (TBP) in the apicomplexan *Cryptosporidium parvum*. Microbiology. 150:1207–1213.

Ozaki J, Takemaru K, Ikegami T, Mishima M, Ueda H, Hirose S, Kabe Y, Handa H, Shirakawa M. 1999. Identification of the core domain and the secondary structure of the transcriptional coactivator MBF1. Genes Cells. 4:415–424.

Pazos F, Helmer-Citterich M, Ausiello G, Valencia A. 1997. Correlated mutations contain information about protein-protein interaction. J Mol Biol. 271:511–523.

Perriere G, Gouy M. 1996. WWW-Query: an on-line retrieval system for biological sequence banks. Biochimie. 78:364–369.

Ramani AK, Marcotte EM. 2003. Exploiting the co-evolution of interacting proteins to discover interaction specificity. J Mol Biol. 327:273–284.

Rizhsky L, Liang H, Mittler R. 2002. The combined effect of drought stress and heat shock on gene expression in tobacco. Plant Physiol. 130:1143–1151.

Saitou N, Nei M. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol. 4:406–425.

Smith ML, Johanson RA, Roger KE, Coleman PD, Slemmon JR. 1998. Identification of a neuronal calmodulin-binding peptide, CAP-19, containing an IQ motif. Brain Res Mol Brain Res. 62:12–24.

Stevens JL, Cantin GT, Wang G, Shevchenko A, Shevchenko A, Berk AJ. 2002. Transcription control by E1A and MAP kinase pathway via Sur2 mediator subunit. Science. 296:755–758.

Taatjes DJ, Marr MT, Tjian R. 2004. Regulatory diversity among metazoan co-activator complexes. Nat Rev Mol Cell Biol. 5:403–410.

Takemaru K, Harashima S, Ueda H, Hirose S. 1998. Yeast coactivator MBF1 mediates GCN4-dependent transcriptional activation. Mol Cell Biol. 18:4971–4976.

Takemaru K, Li FQ, Ueda H, Hirose S. 1997. Multiprotein bridging factor 1 (MBF1) is an evolutionarily conserved transcriptional coactivator that connects a regulatory factor and TATA element-binding protein. Proc Natl Acad Sci USA. 94:7251–7256.

Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG. 1997. The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res. 25:4876–4882.

Zegzouti H, Jones B, Frasse P, Marty C, Maitre B, Latch A, Pech JC, Bouzayen M. 1999. Ethylene-regulated gene expression in tomato fruit: characterization of novel ethylene-responsive and ripening-related genes isolated by differential display. Plant J. 18:589–600.

Yoko Satta, Associate Editor

Accepted March 15, 2007

DOI: 10.1534/genetics.106.069484

Transcriptional Interferences in *cis* Natural Antisense Transcripts of Humans and Mice

Naoki Osato, Yoshiyuki Suzuki, Kazuho Ikeo and Takashi Gojobori[1]

Center for Information Biology and DNA Data Bank of Japan, National Institute of Genetics, Research Organization of Information and Systems, Mishima 411-8540, Japan

Manuscript received December 9, 2006
Accepted for publication March 21, 2007

ABSTRACT

For a significant fraction of mRNAs, their expression is regulated by other RNAs, including *cis* natural antisense transcripts (*cis*-NATs) that are complementary mRNAs transcribed from opposite strands of DNA at the same genomic locus. The regulatory mechanism of mRNA expression by *cis*-NATs is unknown, although a few possible explanations have been proposed. To understand this regulatory mechanism, we conducted a large-scale analysis of the currently available data and examined how the overlapping arrangements of *cis*-NATs affect their expression level. Here, we show that for both human and mouse the expression level of *cis*-NATs decreases as the length of the overlapping region increases. In particular, the proportions of the highly expressed *cis*-NATs in all *cis*-NATs examined were ~36 and 47% for human and mouse, respectively, when the overlapping region was <200 bp. However, both proportions decreased to virtually zero when the overlapping regions were >2000 bp in length. Moreover, the distribution of the expression level of *cis*-NATs changes according to different types of the overlapping pattern of *cis*-NATs in the genome. These results are consistent with the transcriptional collision model for the regulatory mechanism of gene expression by *cis*-NATs.

BIOLOGICAL processes such as development, metabolism, and response to external stimuli are conducted by the cooperative activities of many genes. To understand a biological process, it is essential to understand the regulatory network of genes composing the biological process. After genome sequences had been determined, attempts to reveal regulatory networks of genes were started (Wyrick and Young 2002; Encode Project Consortium 2004; Carninci *et al.* 2005; Levine and Davidson 2005). Regulation of gene expression can be conducted mainly by proteins such as transcription factors. However, it has been found that ~20–30% of mammalian transcripts are targets of microRNAs, which bind to complementary mRNAs and inhibit their activation (Krek *et al.* 2005; Lewis *et al.* 2005; Stark *et al.* 2005; Carthew 2006). This suggests that the regulation of gene expression by RNAs is more ubiquitous and important than we thought (Mattick 2001, 2004).

Cis natural antisense transcripts (*cis*-NATs) are composed of a pair of mRNAs that are transcribed from the opposite strands of DNA at the same genomic locus. The antisense mRNA regulates the expression level of the sense mRNA in a pair. As a result, *cis*-NATs affect the developmental processes such as neural, eye, and tooth formation (Potter and Branford 1998; Korneev *et al.* 1999; Alfano *et al.* 2005; Coudert *et al.* 2005; Korneev and O'Shea 2005), and various molecular functions such as X-inactivation, genomic imprinting, DNA methylation, RNA editing, and alternative splicing (Munroe and Lazar 1991; Kumar and Carmichael 1997; Moore *et al.* 1997; Lee *et al.* 1999; Tufarelli *et al.* 2003).

Although the number of experimentally verified *cis*-NATs was ~40, >2000 *cis*-NATs were predicted in the analyses of the genomic and cDNA sequences in humans and mice (Lehner *et al.* 2002; Shendure and Church 2002; Kiyosawa *et al.* 2003; Yelin *et al.* 2003; Katayama *et al.* 2005). Recently, an analysis of a human oligo microarray showed that as much as 60% of surveyed loci on human chromosome 10 were predicted to encode *cis*-NATs (Cheng *et al.* 2005). Other chromosomes were also expected to encode as many *cis*-NATs. In addition, the presence of *cis*-NATs has been predicted for other eukaryotes as well as prokaryotes (Wagner and Simons 1994; Vanhee-Brossollet and Vaquero 1998; Makalowska *et al.* 2005). These observations implied that the regulation of gene expression by *cis*-NATs would occur more frequently than previously considered. Regulation of gene expression by RNAs may be evolutionarily advantageous, because it regulates gene expression quickly and saves energy and time in synthesizing proteins. Chen *et al.* (2005b) showed that *cis*-NATs were encoded in genes with shorter intron sequences than other mRNAs.

[1] *Corresponding author:* Center for Information Biology and DNA Data Bank of Japan, National Institutes of Genetics, Yata 1111, Mishima, Shizuoka 411-8540, Japan. E-mail: tgojobor@genes.nig.ac.jp

Although a large number of *cis*-NATs have been predicted from various species, the regulatory mechanisms of gene expression by *cis*-NATs remain unclear. To understand the regulatory mechanisms, it will be crucial to know the features of *cis*-NATs that are important in the regulation of gene expression. Thus far, three models have been proposed for the regulation of gene expression by *cis*-NATs (Lavorgna *et al.* 2004).

The first model asserts that *cis*-NATs form a double strand through their complementary sequences, which leads to the inhibition of the function of mRNAs, including protein synthesis. In this model, it is expected that *cis*-NATs are overlapped in at least 6- to 8-bp regions to form stable double strands of RNA (Lai 2002; Lewis *et al.* 2005).

The second model involves epigenetic regulations such as the methylation of promoters and the conversion of the chromosome structure (Wutz *et al.* 1997; Reik and Walter 2001; Tufarelli *et al.* 2003). Through unknown mechanisms, the antisense mRNAs methylate promoters of sense mRNAs and inhibit the transcription of sense mRNAs. In addition, the antisense mRNAs convert the chromosomal structures in which *cis*-NATs are located and regulate the expression of the sense mRNAs. For the model of epigenetic regulations, the features of *cis*-NATs that are essential for the regulations are unclear.

The third model is transcriptional collisions. In the transcription of *cis*-NATs, RNA polymerases bind to the promoters of genes encoding sense and antisense mRNAs and synthesize mRNAs, moving toward the 3′-end of the genes. RNA polymerases clash in the overlapping region and inhibit their transcription (Figure 4). This model was implied from the analyses of the expression levels of *cis*-NATs in yeast (Peterson and Myers 1993; Puig *et al.* 1999; Prescott and Proudfoot 2002). In the analyses, the expression level of *cis*-NATs decreased as the length of the overlapping region of the *cis*-NATs increased. Moreover, the expression level of adjacent transcripts in tandem on the same strand of the yeast genome decreased when the terminator of the upstream transcript was removed. This suggested that RNA polymerases did not stop at the terminator of the upstream transcript and affected the transcription of the downstream transcript. Recently, the collision of *Escherichia coli* RNA polymerases was observed by atomic force microscopy (Crampton *et al.* 2006). This observation showed that RNA polymerases do not pass each other or displace one another, but instead stall against each other.

Here, we report the effects of length and pattern of the overlapping regions of *cis*-NATs for humans and mice on their expression level. Moreover, human and mouse adjacent transcripts, in a particular position, affect their expression level. These results are consistent with the transcriptional collision model, implying that the regulation of the expression of *cis*-NATs by transcriptional collisions is common among species.

MATERIALS AND METHODS

Analysis of the expression level of human *cis*-NATs: To predict *cis*-NATs from human cDNA sequences, we collected a total of 46,675 human cDNA sequences, which consisted of 39,530 human Ensembl cDNA sequences (February 2006) (Hubbard *et al.* 2005), 6501 human Ensembl non-protein-coding sequences (February 2006), and 644 non-protein-coding sequences in RNAdb (Pang *et al.* 2005) that were mapped to the human genome by BLAT software (Kent 2002). We predicted 8964 *cis*-NATs, which had an at least 1-bp-long overlapping region, on the basis of their genomic location using in-house Perl scripts. Redundant *cis*-NATs were merged into the same group when *cis*-NATs overlapped in an at least 1-bp-long region in the genome. The number of the groups of *cis*-NATs was 2496. To examine the expression levels of human *cis*-NATs, we employed 15.5 million *Nla*III human serial analysis of gene expression (SAGE) tags (November 2005) of all tissues in the NCBI SAGEmap database (Lash *et al.* 2000). Human SAGE tags were searched against a total of 46,675 human cDNA sequences according to the protocol for SAGE (Velculescu *et al.* 1995) using in-house Perl scripts. When a SAGE tag matched more than one transcript, these transcripts were removed from further analyses. As a result, among the 46,675 human cDNA sequences, 28,009 had unique SAGE tag assignments, and among the 2496 groups of human *cis*-NATs, 728 groups had unique SAGE tag assignments to all transcripts in each group. Some SAGE tags are supposed to be assigned to overlapping regions of *cis*-NATs. Although cDNA microarrays cannot distinguish the expression of mRNAs encoded in plus and minus strands of the genome in the same locus, SAGE tags are strand specific. Therefore, even though SAGE tags are produced from the overlapping regions of *cis*-NATs, the expression levels of sense and antisense mRNAs of *cis*-NATs can be measured separately.

To examine the expression levels of the human *cis*-NATs, we compared the expression level of human *cis*-NATs with that of other human transcripts (*i.e.*, human transcripts excluding *cis*-NATs, pseudogenes, and non-protein-coding mRNAs). We calculated the ratio of the expression level of a human *cis*-NAT to that of the other human transcripts. However, the expression levels of transcripts are known to be affected by the overall length of the transcripts (Castillo-Davis *et al.* 2002). Thus, we compensated the expression level of *cis*-NATs according to their overall length. The expression level of a human *cis*-NAT was compared with that of the other human transcripts with almost the same overall length of the *cis*-NAT. We removed human *cis*-NATs, pseudogenes, and non-protein-coding mRNAs from human cDNA sequences and selected 2000 human transcripts of which overall length in the genome was close to the overall length of a human *cis*-NAT in the genome. We calculated the median of the expression levels of the selected 2000 human transcripts (supplemental Figure 1A at http://www.genetics.org/supplemental/). The median of the expression levels of the selected human transcripts was defined as a ratio of 1.0, and then the ratio of the expression level of a *cis*-NAT was calculated as follows:

$$R_{\mathit{cis}\text{-NAT}} = T_{\mathit{cis}\text{-NAT}} / T_{(\text{all}-\mathit{cis}\text{-NAT}-\text{pseudo}-\text{noncoding})};$$

where $R_{\mathit{cis}\text{-NAT}}$ is ratio of the expression level of a human *cis*-NAT to that of other human protein-coding transcripts, $T_{\mathit{cis}\text{-NAT}}$ is the expression level of a human *cis*-NAT, and $T_{(\text{all}-\mathit{cis}\text{-NAT}-\text{pseudo}-\text{noncoding})}$ is the median of the expression level of other human transcripts (*i.e.*, human transcripts excluding *cis*-NATs, pseudogenes, and non-protein-coding mRNAs) with almost the same overall length of the *cis*-NAT.

Sense mRNAs of *cis*-NATs are located not only in the plus or the minus strand of the genome, but also in both strands of the genome. Some of the sense mRNAs encode proteins and others encode non-protein-coding mRNAs that may have some biological function such as the regulation of mRNA translation and stability (Mattick and Makunin 2006). The antisense mRNA in each group of *cis*-NATs is known to decrease the expression level of the sense mRNA in the group and to inhibit the activation of the sense mRNA (Wagner and Simons 1994; Kumar and Carmichael 1998; Vanhee-Brossollet and Vaquero 1998). Therefore, we recognized the *cis*-NAT with the lowest expression level in each group as the sense mRNA in the group. In this study, we used the expression level of the sense mRNA in each group to examine the expression levels of *cis*-NATs according to the overlapping arrangements in the genome. However, there is an assumption that under some regulatory mechanisms of *cis*-NATs, such as RNA masking and a double-stranded RNA-dependent mechanism, the expression level of a sense mRNA may not be the lowest in the group of *cis*-NATs and an inverse correlation of the expression levels may not be found between a sense and an antisense mRNA (Chen *et al.* 2005a; Katayama *et al.* 2005; Lapidot and Pilpel 2006). Therefore, we also examined the expression level of *cis*-NATs randomly selected (*i.e.*, selected in a non-expression-level-dependent manner) (see supplemental material at http://www.genetics.org/supplemental/).

We examined the expression levels of *cis*-NATs as the overlapping regions increased in length. When *cis*-NATs included more than two transcripts (*i.e.*, sense and antisense transcripts on both strands of the genome), the length of the overlapping region was defined as the distance between the farthest upstream and the farthest downstream genomic locations of the overlapping regions of *cis*-NATs in a group.

Analysis of the expression level of mouse *cis*-NATs: To predict the mouse *cis*-NATs, we collected a total of 35,486 mouse cDNA sequences, which consisted of 33,252 mouse Ensembl cDNA sequences (April 2006), 1752 mouse Ensembl non-protein-coding sequences (April 2006), and 482 non-protein-coding sequences in RNAdb that were mapped to the mouse genome using BLAT. We predicted 5491 *cis*-NATs from the mouse cDNA sequences and redundant *cis*-NATs were merged into the same group. There were 1868 groups of mouse *cis*-NATs. We examined the expression levels of mouse *cis*-NATs using 3.6 million *Nla*III mouse SAGE tags of all tissues in the NCBI SAGEmap database (November 2005) to compare them with the expression levels of human *cis*-NATs. Among the 35,486 mouse cDNA sequences, 21,982 had unique SAGE tag assignments, and among the 1868 groups of mouse *cis*-NATs, 704 groups had unique SAGE tag assignments to all transcripts, including alternative forms in each group. We examined the expression levels of mouse *cis*-NATs in the same way we examined those of humans.

Comparison of *cis*-NATs of humans and mice at the nucleotide level: To find *cis*-NATs conserved between humans and mice, we compared 46,675 human cDNA sequences with 35,486 mouse cDNA sequences and vice versa using the all-against-all FASTA (Pearson and Lipman 1988) procedure. An expected (E)-value cutoff of 1.0×10^{-20} was used. We selected the human and mouse cDNA sequences that matched reciprocally with an E-value less than the square root of the lowest E-value as candidates of orthologs. When human *cis*-NATs in both strands of the genome are orthologous to mouse *cis*-NATs in both strands of the genome, we recognized the *cis*-NATs as conserved between human and mouse.

RESULTS

Length of the overlapping region of human *cis*-NATs affects their expression level: To predict *cis*-NATs from human cDNA sequences, we searched 46,675 human cDNA sequences. A total of 8964 *cis*-NATs were predicted and were clustered into 2496 groups, each of which consisted of sense and antisense transcripts as well as their alternative forms. To examine the expression levels of human cDNA sequences, ~15.5 million *Nla*III SAGE tags were collected from all the human tissues available in the NCBI SAGEmap database (November 2005) (Lash *et al.* 2000) and were compared to 46,675 human cDNA sequences. Among the 8964 (2496 groups) *cis*-NATs, 2038 (728 groups) had unique SAGE tag assignments to all transcripts in each group.

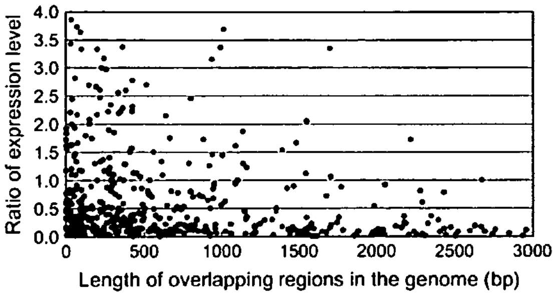


Figure 1.—Distribution of the expression level of human *cis*-NATs as the overlapping region in the genome increased in length. The *x*-axis shows the length of the overlapping exon and intron regions of human *cis*-NATs in the genome. The *y*-axis shows the ratio of the expression level of a human *cis*-NAT to the median of the expression levels of human transcripts of which overall lengths in the genome are close to the overall length of the human *cis*-NAT in the genome. The median of the expression levels of human transcripts is defined as a ratio of 1.0. Because the expression levels of human transcripts are affected by their overall length, we compensated for the expression level of a *cis*-NAT according to its overall length (see materials and methods).

To examine whether the length of overlapping regions in *cis*-NATs affects their expression level, we investigated the relationship between the expression levels of *cis*-NATs and the length of the overlapping region in the genome. It should be noted that Castillo-Davis *et al.* (2002) found that highly expressed transcripts tended to have short introns, implying that the short *cis*-NATs may be expressed at a higher level than the long *cis*-NATs. To eliminate the effect of the overall length of a transcript on its expression level, we selected nonoverlapping transcripts whose overall lengths in the genome were almost the same as that of a *cis*-NAT in the genome and then compared the expression level of the *cis*-NATs with that of the selected transcripts (see materials and methods).

In Figure 1, the *y*-axis represents the ratio of the expression level of the human *cis*-NATs to that of the selected human transcripts. The proportion of highly expressed *cis*-NATs (ratio >1.0) in all *cis*-NATs examined was 36% when the overlapping region was between 1 and 200 bp. However, the proportion decreased to virtually zero when the overlapping regions were >2000 bp long (chi-square $P < 10^{-15}$). This result suggests that the expression level of *cis*-NATs decreases as the length of the overlapping region increases. In addition, the expression levels of *cis*-NATs may be influenced by other factors such as alternative transcripts of *cis*-NATs, the difference of the expression levels among SAGE tag libraries, GC content bias of SAGE tags (Margulies *et al.* 2001), the experimental methods for the analysis of gene expression, the ways of selecting sense mRNAs, the criteria for the length of the overlapping regions, and a set of human cDNA sequences used in this analysis. However, these factors did not change the overall distribution of the expression levels significantly (see supplemental material at http://www.genetics.org/supplemental/).

Overlapping pattern of human *cis*-NATs affects their expression level: *Cis*-NATs are classified into three types on the basis of overlapping patterns in the genome: head to head, tail to tail, and full overlap (Figure 2). "Full overlap" describes *cis*-NATs where the sense mRNA entirely overlaps within the antisense mRNA. The numbers of head-to-head, tail-to-tail, and full-overlap types of human *cis*-NATs were 254, 476, and 1766 groups of which 126, 230, and 356 groups had unique SAGE tag assignments to all transcripts in each group, respectively. To examine whether the overlapping pattern of *cis*-NATs affects their expression level, we analyzed the expression level of *cis*-NATs of humans according to the overlapping patterns in the genome. Figure 3, A–C, shows the expression levels of human *cis*-NATs in the head-to-head, tail-to-tail, and full-overlap manners, respectively. The highly expressed *cis*-NATs decreased in quantity as the overlapping region increased in length for all types of *cis*-NATs. However, highly expressed *cis*-NATs in head-to-head and full-overlap manners decreased in quantity more than those in a tail-to-tail manner did. When the length of the overlapping region was <600 bp, 26.7% of *cis*-NATs in a head-to-head manner showed high expression (ratio >1.0) and 43.4% of *cis*-NATs in a tail-to-tail manner showed high expression. The proportion of highly expressed *cis*-NATs in a head-to-head manner (26.7%) was 1.6 times smaller than that in a tail-to-tail manner (43.4%) (Mann–Whitney *U*-test: $P < 10^{-2}$). Similarly, when the length of the overlapping region was <600 bp, 24.5% of *cis*-NATs in a full-overlap manner showed high expression. The proportion of highly expressed *cis*-NATs in a full-overlap manner (24.5%) was 1.7 times smaller than that in a tail-to-tail manner (43.4%) (Mann–Whitney *U*-test: $P < 10^{-3}$). Among the 356 *cis*-NATs in a full-overlap manner, 314 were *cis*-NATs where a sense transcript overlapped only in the intron regions of the antisense transcript in the genome. The mRNAs that overlapped in the intron regions showed the same feature of expression.

As many as 1450 human transcripts were found to be located within a distance of <1 kbp in the genome (Adachi and Lieber 2002; Koyanagi *et al.* 2005). To examine whether the expression levels of nearby transcripts decreased, we investigated the expression levels of human nearby transcripts. Figure 3D shows the expression levels of nearby transcripts where the 5′-end of a transcript is near the 5′-end of another transcript in the genome. Here, we call them "nearby transcripts in a head-to-head manner" (Figure 2). When the distance between the 5′-ends of transcripts was < ~50 bp, highly expressed transcripts (a ratio >1.0) were not observed (chi-square $P < 10^{-7}$).

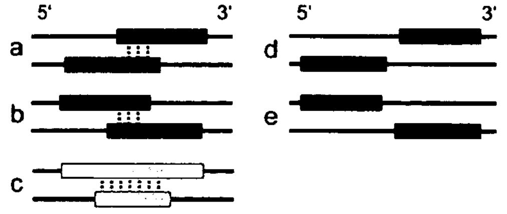


Figure 2.—Classification of *cis*-NATs and nearby transcripts on the basis of their relative positions in the genome. *Cis*-NATs are classified on the basis of their relative positions in the genome: (a) *cis*-NATs in a head-to-head manner (5′-end to 5′-end), (b) those in a tail-to-tail manner (3′-end to 3′-end), and (c) those in a full-overlap manner. Full overlap describes the *cis*-NATs where a transcript on a strand of the genome is overlapped by the entire length of the other transcript on the opposite strand of the genome. (d) Nearby transcripts in a head-to-head manner where the 5′-end of a transcript is near the 5′-end of another transcript in the genome. (e) Nearby transcripts in a tail-to-tail manner where the 3′-end of a transcript is near the 3′-end of another transcript in the genome.

Figure 3E shows the expression levels of nearby transcripts where the 3′-end of a transcript is near the 3′-end of another transcript in the genome. Here, we call them "nearby transcripts in a tail-to-tail manner." Contrary to nearby transcripts in a head-to-head manner, the expression levels of nearby transcripts in a tail-to-tail manner did not change, regardless of the distance of the nearby transcripts in the genome (chi-square $P = 0.7$).

There is a possibility that the length of some human transcripts registered in a database such as the Ensembl database may be shorter than natural transcripts (Makalowska *et al.* 2005). Nearby transcripts found in the Ensembl database may, in fact, overlap in the genome, such that the expression levels of such artificial nearby transcripts seemed to decrease. However, almost all nearby transcripts in a head-to-head manner were expressed at a low level when the distance of the transcripts was < ~50 bp. Therefore, the decrease in the expression level of nearby transcripts will be a natural phenomenon.

Overlapping arrangements of mouse *cis*-NATs affect their expression levels: *Cis*-NATs have been predicted for various species (Wagner and Simons 1994; Vanhee-Brossollet and Vaquero 1998; Wagner and Flardh 2002; Makalowska *et al.* 2005). If the regulatory

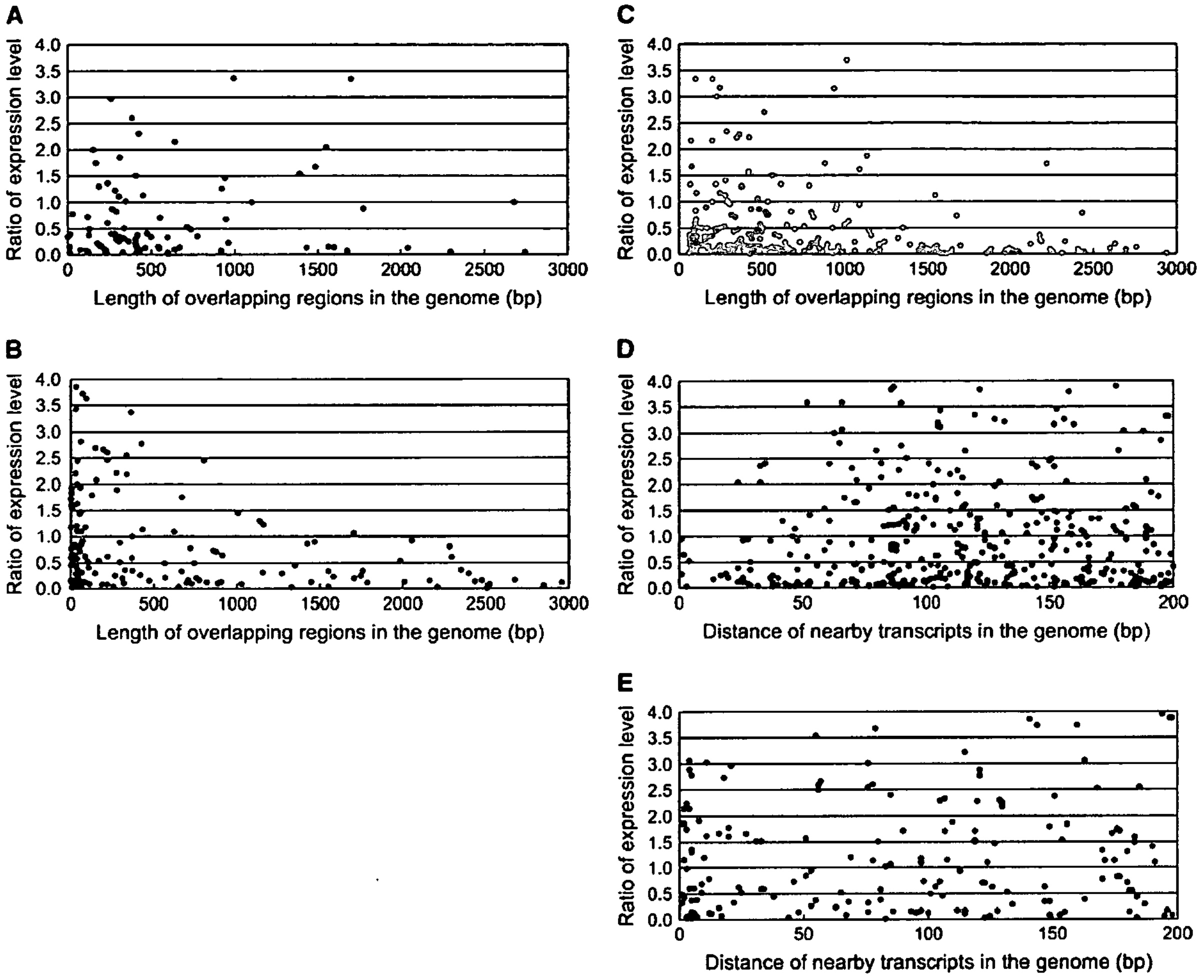


FIGURE 3.—Distribution of the expression level of human *cis*-NATs according to overlapping patterns in the genome. The *x*-axis shows the length of the overlapping exon and intron regions of human *cis*-NATs in the genome. The *y*-axis shows the ratio of the expression level of a *cis*-NAT to the median of the expression levels of human transcripts whose overall lengths in the genome are close to those of the human *cis*-NAT in the genome. The median of the expression levels of human transcripts is defined as a ratio of 1.0. Because the expression levels of human transcripts are affected by their overall length, we compensated the expression level of a *cis*-NAT according to its overall length (see MATERIALS AND METHODS). (A) *cis*-NATs in a head-to-head manner, (B) those in a tail-to-tail manner, (C) those in a full-overlap manner, (D) nearby transcripts in a head-to-head manner, and (E) nearby transcripts in a tail-to-tail manner.

mechanisms of *cis*-NATs in gene expression are conserved among species, *cis*-NATs of another species are expected to show similar effects on their expression levels. To evaluate whether the relationship between the overlapping arrangements and the expression levels of *cis*-NATs is conserved among species, we examined the expression levels of mouse *cis*-NATs. Almost the same number of *cis*-NATs (1771 groups) as that of humans was found in mouse cDNA sequences (KIYOSAWA *et al.* 2003; YELIN *et al.* 2003). Mouse *cis*-NATs were compared to 3.6 million *Nla*III mouse SAGE tags in the NCBI SAGEmap database. Although the number of mouse SAGE tags and the number of mouse *cis*-NATs (705 groups) assigned to unique SAGE tags was smaller than for humans, the distribution of the expression levels of mouse *cis*-NATs showed the same features as that of humans (supplemental Figure 2A at http://www.genetics.org/supplemental/): highly expressed (ratio >1.0) *cis*-NATs decreased in quantity when the overlapping regions in the genome increased in length. The proportion of highly expressed *cis*-NATs in all *cis*-NATs examined was 47% when the overlapping region was between 1 and 200 bp, and the proportion decreased virtually to zero when the overlapping regions were >2000 bp long (chi-square $P < 10^{-15}$).

For overlapping patterns in the genome, the distribution of the expression levels in mice changed in the same way as in humans (Mann–Whitney *U*-test: $P = 0.52$ between human and mouse *cis*-NATs in a head-to-head manner, $P = 0.92$ between those in a tail-to-tail manner, and $P = 0.17$ between those in a full-overlap manner) (supplemental Figure 2, B–D, at http://www.genetics.org/

supplemental/). When the length of the overlapping region was <600 bp, 27.6% of *cis*-NATs in a head-to-head manner showed high expression (ratio >1.0) and 44.1% of *cis*-NATs in a tail-to-tail manner showed high expression. The proportion of highly expressed *cis*-NATs in a head-to-head manner (27.6%) was 1.6 times smaller than that of those in a tail-to-tail manner (44.1%) (Mann–Whitney *U*-test: $P < 10^{-2}$). Similarly, 25.9% of *cis*-NATs in a full-overlap manner showed high expression (ratio >1.0) and the proportion of highly expressed *cis*-NATs in a full-overlap manner (25.9%) was 1.7 times smaller than that in a tail-to-tail manner (44.1%) (Mann–Whitney *U*-test: $P < 0.05$). With nearby transcripts, when the distance between nearby transcripts in a head-to-head manner was < ~50 bp, highly expressed transcripts were not observed as found in humans (chi-square $P < 10^{-5}$) (supplemental Figure 2E at http://www.genetics.org/supplemental/). The expression levels of nearby transcripts in a tail-to-tail manner did not change, regardless of the distance of the nearby transcripts in the genome (chi-square $P = 0.9$) (supplemental Figure 2F at http://www.genetics.org/supplemental/). These results suggest that the expression levels of mouse *cis*-NATs are affected by the overlapping arrangements in the genome in the same way as those of humans. This implies that the regulatory mechanisms of *cis*-NATs in gene expression are conserved between humans and mice.

However, there was a possibility that *cis*-NATs showed a similar distribution of the expression level between humans and mice because most human and mouse *cis*-NATs were orthologous (Liao and Zhang 2006). To address this possibility, we compared the distribution of the expression levels of *cis*-NATs that are not conserved between human and mouse. First, we compared human and mouse cDNA sequences by using FASTA (Pearson and Lipman 1988) and found that only 329 groups (11.9%) of human *cis*-NATs were conserved in mouse *cis*-NATs, although 34,670 (76.0%) of human cDNA sequences were conserved in mice. Human and mouse *cis*-NATs were supposed to be highly divergent in terms of cDNA sequences (Veeramachaneni *et al.* 2004; Makalowska *et al.* 2005). We removed the 329 groups of *cis*-NATs from 2765 groups of human and 1704 groups of mouse *cis*-NATs, which left 710 groups of human and 605 groups of mouse *cis*-NATs with SAGE tags assigned to all transcripts in each group. We examined the expression level of the human and mouse *cis*-NATs according to the length and the pattern of the overlapping region in the genome. They showed almost the same distribution of the expression level as those including *cis*-NATs conserved between humans and mice (Mann–Whitney *U*-test: $P = 0.22$ and $P = 0.88$ for humans and mice, respectively). These results suggested that the similarity of the distribution of the expression level of human and mouse *cis*-NATs was not due to the conservation of the cDNA sequences of the *cis*-NATs.

DISCUSSION

We found that the expression level of *cis*-NATs changed according to the overlapping arrangements of *cis*-NATs in the human and mouse genomes. The expression level of *cis*-NATs decreased when the overlapping regions increased in length. Moreover, the overlapping pattern of *cis*-NATs affects their expression level. Nearby transcripts in a particular position decreased their expression levels.

Here, we examined the expression level of *cis*-NATs using SAGE tags and oligonucleotide arrays in public databases. We obtained the same distribution of the expression level of *cis*-NATs at least in the same tissue or cell such as fetal brain, embryonic stem cell, and liver (supplemental material and supplemental Figures 3 and 4 at http://www.genetics.org/supplemental/). However, all SAGE libraries and the expression data of oligonucleotide arrays were produced from cells and tissues of humans and mice, not from a single cell. In addition, some *cis*-NATs may be expressed only in some developmental stages or at a specific time. Therefore, some *cis*-NATs may not be expressed concurrently in the same single cell.

Thus far, three models have been proposed for the regulation of gene expression by *cis*-NATs. Forming double strands of *cis*-NATs requires a minimum 6- to 8-bp overlapping region of *cis*-NATs. We found that the expression level of human and mouse *cis*-NATs decreased consecutively as the length of the overlapping region increased. However, currently there is no report that this result is brought about by this model. In addition, this model does not intend to explain the change of the expression levels of nearby transcripts. Epigenetic regulations are also considered to be involved in the regulation of *cis*-NATs in gene expression. However, currently there is no report that by this model the expression level of human and mouse *cis*-NATs decreases consecutively as the length of the overlapping region increases.

Our results are consistent with the transcriptional collision model that has been proposed following the analyses of adjacent transcripts and *cis*-NATs in yeast (Puig *et al.* 1999; Prescott and Proudfoot 2002) and the observation of the collision of RNA polymerases by atomic force microscopy (Crampton *et al.* 2006). RNA polymerases bind to the upstream regions of genes and synthesize mRNAs, moving toward the 3′-ends of the genes. When opposite genomic strands in the same locus encode complementary mRNAs like *cis*-NATs, an RNA polymerase bound on a strand of the genome collides with the RNA polymerase bound on the opposite strand during the transcription of both strands of mRNAs (Figure 4). This leads to the inhibition of transcription. From this model, the frequency of the collisions of RNA polymerases is expected to increase when the overlapping regions increase in length. Moreover, overlapping patterns in the genome would affect the

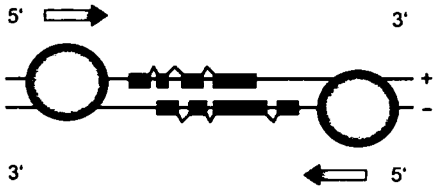


FIGURE 4.—Transcriptional collision model When *cis*-NATs are transcribed by RNA polymerases, RNA polymerases bind to the upstream region of a gene encoding a sense mRNA and synthesize the complementary mRNA, moving to the 3'-end of the gene. Similarly, RNA polymerases that are bound to the upstream region of a gene encoding an antisense mRNA move to the 3'-end of the gene. Then, RNA polymerases collide with each other in the overlapping region of the genes, thereby inhibiting the transcription.

frequency of the collisions of RNA polymerases. In the case of *cis*-NATs in a head-to-head manner, the 5'-ends of mRNAs are the start position for the transcription of mRNAs. Overlapping at the 5'-end would inhibit the initiation of transcription and decrease the expression level of the *cis*-NATs (Figure 3A). Contrary to *cis*-NATs in a head-to-head manner, overlapping at the 3'-end would not decrease the expression level significantly when the overlapping region is short (Figure 3B). In the case of *cis*-NATs in a full-overlap manner, both the 5'- and 3'-ends of a transcript are overlapped. This would decrease the expression level, even when the length of overlapping regions is short (Figure 3C). For nearby transcripts, highly expressed nearby transcripts in a head-to-head manner decreased in quantity when the distance between the nearby transcripts was <50 bp (Figure 3D). However, the level of highly expressed nearby transcripts in a tail-to-tail manner did not decrease (Figure 3E). These results would occur if the start or end positions of transcripts were close to each other. In the initiation of transcription, RNA polymerases bind to the start position of transcription of mRNAs and cover the region between 55 bp upstream and 20 bp downstream (−55 to 20) of the start position (KORZHEVA *et al.* 2000; LEE and YOUNG 2000; MURAKAMI *et al.* 2002). Therefore, these findings suggest that nearby transcripts in a head-to-head manner inhibited the binding of RNA polymerases to the upstream regions of the transcripts, when the distance between the nearby transcripts in a head-to-head manner was <50 bp. In addition, the model of transcriptional collisions for *cis*-NATs may explain an observation that experiments of Northern hybridization showed smear bands of mRNAs at the genomic regions where *cis*-NATs were located (KIYOSAWA *et al.* 2005). This implies that various lengths of single-stranded mRNAs may be produced by the inhibition of the transcription and unusual movements of RNA polymerases.

Among the 2462 groups of human *cis*-NATs, 874 groups did not include alternative forms of sense and antisense mRNAs, and among the 874 groups, 542 (62%) groups consisted of *cis*-NATs where a sense mRNA overlapped only the intron regions of the gene encoding the antisense mRNA in the genome. As shown in Figures 1 and 3, the expression level of *cis*-NATs overlapping the intron regions also decreased as the length of the overlapping region increased. As for the regulatory mechanisms of gene expression by *cis*-NATs overlapping the intron regions, a double-stranded RNA-dependent mechanism would be difficult to use in explaining the decrease of the expression level of the *cis*-NATs, because they cannot form double strands of mRNAs after transcription and pre-mRNA splicing of mRNAs. Although double strands of mRNAs may be formed before pre-mRNA splicing after transcription, it is unclear whether it occurs. In the meantime, transcriptional collisions reasonably explain that *cis*-NATs overlapping the intron regions affected the expression of the *cis*-NATs.

Our observations are consistent with the transcriptional collision model. However, this does not mean that they exclude other regulatory mechanisms from the regulation of *cis*-NATs in gene expression. In addition to the regulation by general transcription factors in gene expression, *cis*-NATs employ several regulatory mechanisms, including transcriptional collisions (LAVORGNA *et al.* 2004). Our findings will be useful for the examination and understanding of the regulatory mechanisms of *cis*-NATs in gene expression and furthermore will help in elucidating the regulatory network of genes and their evolution.

We are grateful to members of the Laboratory for DNA Data Analysis at the National Institute of Genetics for discussion and comments on the manuscript. This work was supported by a research grant from the Institute for Bioinformatics Research and Development, Japan Science and Technology Agency, and by a grant of the Genome Network Project from the Ministry of Education, Culture, Sports, Science and Technology, Japan.

LITERATURE CITED

ADACHI, N., and M. R. LIEBER, 2002 Bidirectional gene organization: a common architectural feature of the human genome. Cell **109:** 807–809.

ALFANO, G., C. VITIELLO, C. CACCIOPPOLI, T. CARAMICO, A. CAROLA *et al.*, 2005 Natural antisense transcripts associated with genes involved in eye development. Hum. Mol. Genet. **14:** 913–923.

CARNINCI, P., T. KASUKAWA, S. KATAYAMA, J. GOUGH, M. C. FRITH *et al.*, 2005 The transcriptional landscape of the mammalian genome. Science **309:** 1559–1563.

CARTHEW, R. W., 2006 Gene regulation by microRNAs. Curr. Opin. Genet. Dev. **16:** 203–208.

CASTILLO-DAVIS, C. I., S. L. MEKHEDOV, D. L. HARTL, E. V. KOONIN and F. A. KONDRASHOV, 2002 Selection for short introns in highly expressed genes. Nat. Genet. **31:** 415–418.

CHEN, J., M. SUN, L. D. HURST, G. G. CARMICHAEL and J. D. ROWLEY, 2005a Genome-wide analysis of coordinate expression and evolution of human *cis*-encoded sense-antisense transcripts. Trends Genet. **21:** 326–329.

CHEN, J., M. SUN, L. D. HURST, G. G. CARMICHAEL and J. D. ROWLEY, 2005b Human antisense genes have unusually short introns: evidence for selection for rapid transcription. Trends Genet. **21:** 203–207.

Cheng, J., P. Kapranov, J. Drenkow, S. Dike, S. Brubaker *et al.*, 2005 Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution. Science **308:** 1149–1154.

Coudert, A. E., L. Pibouin, B. Vi-Fane, B. L. Thomas, M. Macdougall *et al.*, 2005 Expression and regulation of the Msx1 natural antisense transcript during development. Nucleic Acids Res. **33:** 5208–5218.

Crampton, N., W. A. Bonass, J. Kirkham, C. Rivetti and N. H. Thomson, 2006 Collision events between RNA polymerases in convergent transcription studied by atomic force microscopy. Nucleic Acids Res. **34:** 5416–5425.

Encode Project Consortium, 2004 The ENCODE (ENCyclopedia Of DNA Elements) Project. Science **306:** 636–640.

Hubbard, T., D. Andrews, M. Caccamo, G. Cameron, Y. Chen *et al.*, 2005 Ensembl 2005. Nucleic Acids Res. **33:** D447–D453.

Katayama, S., Y. Tomaru, T. Kasukawa, K. Waki, M. Nakanishi *et al.*, 2005 Antisense transcription in the mammalian transcriptome. Science **309:** 1564–1566.

Kent, W. J., 2002 BLAT: the BLAST-like alignment tool. Genome Res. **12:** 656–664.

Kiyosawa, H., I. Yamanaka, N. Osato, S. Kondo and Y. Hayashizaki, 2003 Antisense transcripts with FANTOM2 clone set and their implications for gene regulation. Genome Res. **13:** 1324–1334.

Kiyosawa, H., N. Mise, S. Iwase, Y. Hayashizaki and K. Abe, 2005 Disclosing hidden transcripts: mouse natural sense-antisense transcripts tend to be poly(A) negative and nuclear localized. Genome Res. **15:** 463–474.

Korneev, S., and M. O'Shea, 2005 Natural antisense RNAs in the nervous system. Rev. Neurosci. **16:** 213–222.

Korneev, S. A., J. H. Park and M. O'Shea, 1999 Neuronal expression of neural nitric oxide synthase (nNOS) protein is suppressed by an antisense RNA transcribed from an NOS pseudogene. J. Neurosci. **19:** 7711–7720.

Korzheva, N., A. Mustaev, M. Kozlov, A. Malhotra, V. Nikiforov *et al.*, 2000 A structural model of transcription elongation. Science **289:** 619–625.

Koyanagi, K. O., M. Hagiwara, T. Itoh, T. Gojobori and T. Imanishi, 2005 Comparative genomics of bidirectional gene pairs and its implications for the evolution of a transcriptional regulation system. Gene **353:** 169–176.

Krek, A., D. Grun, M. N. Poy, R. Wolf, L. Rosenberg *et al.*, 2005 Combinatorial microRNA target predictions. Nat. Genet. **37:** 495–500.

Kumar, M., and G. G. Carmichael, 1997 Nuclear antisense RNA induces extensive adenosine modifications and nuclear retention of target transcripts. Proc. Natl. Acad. Sci. USA **94:** 3542–3547.

Kumar, M., and G. G. Carmichael, 1998 Antisense RNA: function and fate of duplex RNA in cells of higher eukaryotes. Microbiol. Mol. Biol. Rev. **62:** 1415–1434.

Lai, E. C., 2002 Micro RNAs are complementary to 3′ UTR sequence motifs that mediate negative post-transcriptional regulation. Nat. Genet. **30:** 363–364.

Lapidot, M., and Y. Pilpel, 2006 Genome-wide natural antisense transcription: coupling its regulation to its different regulatory mechanisms. EMBO Rep. **7:** 1216–1222.

Lash, A. E., C. M. Tolstoshev, L. Wagner, G. D. Schuler, R. L. Strausberg *et al.*, 2000 SAGEmap: a public gene expression resource. Genome Res. **10:** 1051–1060.

Lavorgna, G., D. Dahary, B. Lehner, R. Sorek, C. M. Sanderson *et al.*, 2004 In search of antisense. Trends Biochem. Sci. **29:** 88–94.

Lee, J. T., L. S. Davidow and D. Warshawsky, 1999 Tsix, a gene antisense to Xist at the X-inactivation centre. Nat. Genet. **21:** 400–404.

Lee, T. I., and R. A. Young, 2000 Transcription of eukaryotic protein-coding genes. Annu. Rev. Genet. **34:** 77–137.

Lehner, B., G. Williams, R. D. Campbell and C. M. Sanderson, 2002 Antisense transcripts in the human genome. Trends Genet. **18:** 63–65.

Levine, M., and E. H. Davidson, 2005 Gene regulatory networks for development. Proc. Natl. Acad. Sci. USA **102:** 4936–4942.

Lewis, B. P., C. B. Burge and D. P. Bartel, 2005 Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. Cell **120:** 15–20.

Liao, B. Y., and J. Zhang, 2006 Evolutionary conservation of expression profiles between human and mouse orthologous genes. Mol. Biol. Evol. **23:** 530–540.

Makalowska, I., C. F. Lin and W. Makalowski, 2005 Overlapping genes in vertebrate genomes. Comput. Biol. Chem. **29:** 1–12.

Margulies, E. H., S. L. Kardia and J. W. Innis, 2001 Identification and prevention of a GC content bias in SAGE libraries. Nucleic Acids Res. **29:** E60.

Mattick, J. S., 2001 Non-coding RNAs: the architects of eukaryotic complexity. EMBO Rep. **2:** 986–991.

Mattick, J. S., 2004 RNA regulation: A new genetics? Nat. Rev. Genet. **5:** 316–323.

Mattick, J. S., and I. V. Makunin, 2006 Non-coding RNA. Hum. Mol. Genet. **15** (Spec. no. 1): R17–R29.

Moore, T., M. Constancia, M. Zubair, B. Bailleul, R. Feil *et al.*, 1997 Multiple imprinted sense and antisense transcripts, differential methylation and tandem repeats in a putative imprinting control region upstream of mouse Igf2. Proc. Natl. Acad. Sci. USA **94:** 12509–12514.

Munroe, S. H., and M. A. Lazar, 1991 Inhibition of c-erbA mRNA splicing by a naturally occurring antisense RNA. J. Biol. Chem. **266:** 22083–22086.

Murakami, K. S., S. Masuda, E. A. Campbell, O. Muzzin and S. A. Darst, 2002 Structural basis of transcription initiation: an RNA polymerase holoenzyme-DNA complex. Science **296:** 1285–1290.

Pang, K. C., S. Stephen, P. G. Engstrom, K. Tajul-Arifin, W. Chen *et al.*, 2005 RNAdb: a comprehensive mammalian noncoding RNA database. Nucleic Acids Res. **33:** D125–D130.

Pearson, W. R., and D. J. Lipman, 1988 Improved tools for biological sequence comparison. Proc. Natl. Acad. Sci. USA **85:** 2444–2448.

Peterson, J. A., and A. M. Myers, 1993 Functional analysis of mRNA 3′ end formation signals in the convergent and overlapping transcription units of the *S. cerevisiae* genes RHO1 and MRP2. Nucleic Acids Res. **21:** 5500–5508.

Potter, S. S., and W. W. Branford, 1998 Evolutionary conservation and tissue-specific processing of Hoxa 11 antisense transcripts. Mamm. Genome **9:** 799–806.

Prescott, E. M., and N. J. Proudfoot, 2002 Transcriptional collision between convergent genes in budding yeast. Proc. Natl. Acad. Sci. USA **99:** 8796–8801.

Puig, S., J. E. Perez-Ortin and E. Matallana, 1999 Transcriptional and structural study of a region of two convergent overlapping yeast genes. Curr. Microbiol. **39:** 369–373.

Reik, W., and J. Walter, 2001 Genomic imprinting: parental influence on the genome. Nat. Rev. Genet. **2:** 21–32.

Shendure, J., and G. M. Church, 2002 Computational discovery of sense-antisense transcription in the human and mouse genomes. Genome Biol. **3:** RESEARCH0044.

Stark, A., J. Brennecke, N. Bushati, R. B. Russell and S. M. Cohen, 2005 Animal MicroRNAs confer robustness to gene expression and have a significant impact on 3′UTR evolution. Cell **123:** 1133–1146.

Tufarelli, C., J. A. Stanley, D. Garrick, J. A. Sharpe, H. Ayyub *et al.*, 2003 Transcription of antisense RNA leading to gene silencing and methylation as a novel cause of human genetic disease. Nat. Genet. **34:** 157–165.

Vanhee-Brossollet, C., and C. Vaquero, 1998 Do natural antisense transcripts make sense in eukaryotes? Gene **211:** 1–9.

Veeramachaneni, V., W. Makalowski, M. Galdzicki, R. Sood and I. Makalowska, 2004 Mammalian overlapping genes: the comparative perspective. Genome Res. **14:** 280–286.

Velculescu, V. E., L. Zhang, B. Vogelstein and K. W. Kinzler, 1995 Serial analysis of gene expression. Science **270:** 484–487.

Wagner, E. G., and K. Flardh, 2002 Antisense RNAs everywhere? Trends Genet. **18:** 223–226.

Wagner, E. G., and R. W. Simons, 1994 Antisense RNA control in bacteria, phages, and plasmids. Annu. Rev. Microbiol. **48:** 713–742.

Wutz, A., O. W. Smrzka, N. Schweifer, K. Schellander, E. F. Wagner *et al.*, 1997 Imprinted expression of the Igf2r gene depends on an intronic CpG island. Nature **389:** 745–749.

Wyrick, J. J., and R. A. Young, 2002 Deciphering gene expression regulatory networks. Curr. Opin. Genet. Dev. **12:** 130–136.

Yelin, R., D. Dahary, R. Sorek, E. Y. Levanon, O. Goldstein *et al.*, 2003 Widespread occurrence of antisense transcription in the human genome. Nat. Biotechnol. **21:** 379–386.

Communicating editor: S. Yokoyama

D104–D109 Nucleic Acids Research, 2007, Vol. 35, Database issue
doi:10.1093/nar/gkl854

Published online 27 November 2006

H-DBAS: Alternative splicing database of completely sequenced and manually annotated full-length cDNAs based on H-Invitational

Jun-ichi Takeda[1,2], Yutaka Suzuki[3], Mitsuteru Nakao[4,5], Tsuyoshi Kuroda[6], Sumio Sugano[3], Takashi Gojobori[2,7] and Tadashi Imanishi[2,8,*]

[1]Integrated Database Group, Japan Biological Information Research Center, Japan Biological Informatics Consortium, AIST Bio-IT Research Building, Aomi 2-42, Koto-ku, Tokyo 135-0064, Japan, [2]Biological Information Research Center, National Institute of Advanced Industrial Science and Technology, AIST Bio-IT Research Building, Aomi 2-42, Koto-ku, Tokyo 135-0064, Japan, [3]Department of Medical Genome Sciences, Graduate School of Frontier Sciences, the University of Tokyo, 5-1-5 Kashiwanoha, Kashiwa, Chiba 277-8562, Japan, [4]Computational Biology Research Center, National Institute of Advanced Science and Technology, AIST Bio-IT Research Building, Aomi 2-42, Koto-ku, Tokyo 135-0064, Japan, [5]Kazusa DNA Research Institute, 2-6-7 Kazusa-Kamatari, Kisarazu, Chiba 292-0818, Japan, [6]Maze Corporation, TS Building 101, 3-20-2 Hatagaya, Shibuya-ku, Tokyo 151-0072, Japan, [7]Center for Information Biology and DDBJ, National Institute of Genetics, 1111 Yata, Mishima, Shizuoka 411-8540, Japan and [8]Graduate School of Information Science and Technology, Hokkaido University, North 14, West 9, Kita-ku, Sapporo, Hokkaido 060-0814, Japan

Received August 15, 2006; Revised October 9, 2006; Accepted October 10, 2006

ABSTRACT

The Human-transcriptome DataBase for Alternative Splicing (H-DBAS) is a specialized database of alternatively spliced human transcripts. In this database, each of the alternative splicing (AS) variants corresponds to a completely sequenced and carefully annotated human full-length cDNA, one of those collected for the H-Invitational human-transcriptome annotation meeting. H-DBAS contains 38 664 representative alternative splicing variants (RASVs) in 11 744 loci, in total. The data is retrievable by various features of AS, which were annotated according to manual annotations, such as by patterns of ASs, consequently invoked alternations in the encoded amino acids and affected protein motifs, GO terms, predicted subcellular localization signals and transmembrane domains. The database also records recently identified very complex patterns of AS, in which two distinct genes seemed to be bridged, nested or degenerated (multiple CDS): in all three cases, completely unrelated proteins are encoded by a single locus. By using AS Viewer, each AS event can be analyzed in the context of full-length cDNAs, enabling the user's empirical understanding of the relation between AS event and the consequent alternations in the encoded amino acid sequences together with various kinds of affected protein motifs. H-DBAS is accessible at http://jbirc.jbic.or.jp/h-dbas/.

INTRODUCTION

Alternative splicing (AS) is a phenomenon in which various combinations of exons are integrated into different types of transcripts. By utilizing AS, diverse transcripts can be produced. Although it might not be always true that all the variants are translated, this mechanism at least enables a single locus to encode functionally divergent proteins. Actual abundant cases have been reported for such diversification of the gene functions mediated by AS, in which the binding site of a growth factor receptor or an activation site of transcription factor are modified. Especially in mammals, use of AS is widespread [it is reported that 40–60% of entire human genes have AS variants (1)] and is supposed to provide a molecular basis for highly fabricated systems, such as immune systems and neural networks.

Because of the growing interests in AS, a number of databases were launched, such as ASD [http://www.ebi.ac.uk/asd/; (2)] and ASAP [http://www.bioinformatics.ucla.edu/ASAP/; (3)]. However, most of these preexisting AS databases are still incomplete in a sense that they are mainly based on the fragmented information of partially and imprecisely sequenced cDNAs (ESTs) or computationally divided information of the exons. In order to elucidate the functional

*To whom correspondence should be addressed. Tel: +81 3 3599 8800; Fax: +81 3 3599 8801; Email: imanishi@jbirc.aist.go.jp