Biochemical and Biophysical Research Communications **272,** 36–40 (2000)
doi:10.1006/bbrc.2000.2731, available online at http://www.idealibrary.com on IDEAL®

Association of Functional Microsatellites in the Human Type I Collagen α2 Chain (COL1A2) Gene with Systemic Sclerosis

Ryu-Ichiro Hata,*[,1] Jun Akai,* Akinori Kimura,* Osamu Ishikawa,† Masataka Kuwana,‡ and Hiroshi Shinkai§

**Department of Molecular Pathogenesis, Division of Adult Diseases, and Etiology and Pathogenesis Research Unit, Medical Research Institute, Tokyo Medical and Dental University, Tokyo, Japan; †Department of Dermatology, School of Medicine, Gunma University, Maebashi, Gunma, Japan; ‡Division of Cellular Signaling, Institute for Advanced Medicine, Keio University School of Medicine, Tokyo, Japan; and §Department of Dermatology, School of Medicine, Chiba University, Chiba, Japan*

Received March 27, 2000

Systemic sclerosis (SSc) or scleroderma is a generalized disorder of connective tissue. The etiology is poorly understood; however, both genetic and environmental factors have been implicated. To investigate the disease-susceptible gene for SSc, we examined the association of the disease with a gene (COL1A2) for type I collagen, which accumulates excessively in the affected organs. The COL1A2 gene containing a specific combination of the two dinucleotide repeats, repeat-haplotype, is involved in the regulation of gene expression. Homozygotes for a 5′-(CA)13CGCACA(CG)6(CA)8 -(GT)12 -3′ were found with significantly higher frequency ($P = 0.029$, relative risk, RR > 6.93) in SSc patients than in controls, and association was prominent ($P = 0.0042$, RR > 32.0) in the male patients positive for SSc-specific antinuclear antibodies (ANAs). This repeat-haplotype showed the highest stimulative activity for the transcription of the COL1A2 promoter among the reporter gene constructs tested. The results indicate that a portion of the patients having a specific dinucleotide repeat-haplotype homozygously and expressing the ANAs have a significantly higher risk for SSc than those individuals with other combinations of the repeat-haplotypes.

Key Words: **scleroderma; etiology; dinucleotide repeats.**

Systemic sclerosis (scleroderma; SSc) is a generalized disorder of connective tissue characterized by fibrosclerosis and degenerative changes in the skin, synovium, muscles, and certain internal organs, notably the gastrointestinal tract, lung, heart, and kidney (1). Although there are occasionally inflammatory changes in the early stage of the disease, the hallmark of the disease is skin thickening caused by excessive accumulation of connective tissue dominated by type I collagen, which is encoded by COL1A1 and COL1A2 genes. The etiology of SSc is unknown however, both genetic and environmental factors have been implicated. Genetic contributions from the major histocompatibility complex or HLA region have recently been established, especially between SSc-specific autoantibodies and HLA-DQ and DR genes (2, 3).

Other genes contributing to the disease susceptibility have not been identified, but their presence is suggested by data from animal models of SSc (4, 5). SSc rarely occurs in more than one member of a family; however, a number of such kindreds have been reported (6, 7). Furthermore, the pattern of inheritance suggests that SSc may result from multigenic effects similar to those in other autoimmune diseases (3). Genetic mapping of the disease susceptibility genes in SSc is difficult because of its infrequent familial clustering, and potential genetic heterogeneity (3, 7, 8). Recently the association of microsatellite markers near the fibrillin 1 gene of human chromosome 15q with SSc was reported in a Native American population residing in southeastern Oklahoma and having a high prevalence of SSc (8), even though the pathophysiological meaning of the association is not known.

We have been investigating the regulational mechanisms of type I collagen, which is the most abundant

The nucleotide sequence data of the region of COL1A2 gene appeared in the DDBJ, EMBL, and GenBank nucleotide sequence databases under Accession No. AB004317.

[1] To whom correspondence should be addressed at Department of Biochemistry and Molecular Biology, Kanagawa Dental College, 82 Inaoka-cho, Yokosuka, 238-8580, Japan. Fax: +81-468-22-8839. E-mail: ryuhata@kdcnet.ac.jp.

0006-291X/00 $35.00

protein in our body and also the predominant protein deposited in the fibrosclerotic tissue. Production of type I collagen is regulated by multiple steps, among them transcriptional regulation is the most important one (9, 10).

Recently we found two polymorphic dinucleotide repeats (microsatellites) in the presumptive transcriptional regulatory region of the type I collagen $\alpha 2$ chain (COL1A2) gene, which is located on human chromosome 7q21.3-22.1 (11). One was found in the 5′-flanking region of the gene (upstream repeat) and was composed of poly(dC-dA) and poly(dC-dG), whereas the other occurred in the 1st intron (intron repeat), and consisted of poly(dG-dT). These dinucleotide repeats are not just markers of the gene but are also functional, because the co-presence of the two dinucleotide repeats, but not either repeat alone, has an enhancing activity on transcription of the COL1A2 gene as observed from transient transfection experiments of the constructs containing one or both of dinucleotide repeats. In addition, the variety in the number of repetitions is partly responsible for the difference in the transcription activity of the gene (12). Here we investigated the possible association of certain combinations of the two dinucleotide repeats, or repeat-haplotypes, with SSc to find the genetic background of the disease.

MATERIALS AND METHODS

Patients and controls. Twenty-three male and 70 female SSc patients who satisfied the American College of Rheumatology (ACR; formerly the American Rheumatism Association) preliminary classification criteria for SSc (13) were studied. For controls, we selected 209 healthy volunteers (12). We also investigated 91 patients with sporadic dilated cardiomyopathy (DCM) as non-SSc patients.

Antinuclear antigen assays. Antinuclear antibodies were examined by an indirect immunofluorescence assay using HEp-2 cells as a substrate. Individual antibody titers were further determined by an enzyme-linked immunosorbent assay using recombinant antigens (MBL, Nagoya). In some of the patients the presence of anti-DNA topoisomerase I (anti-topo I) antibody was also identified by a double immunodiffusion method using rabbit thymus as an antigen, and was confirmed by immunoprecipitation of 35S-labeled HeLa cell extracts as described previously (14).

Analysis of polymorphisms in the two dinucleotide repeats. Genomic DNA samples were obtained from peripheral leukocytes by the standard sodium dodecyl sulfate-protease K digestion followed by phenol-chloroform extraction (12). Length polymorphism at the dinucleotide repeat regions was examined by electrophoresis of polymerase chain reaction (PCR) fragments on 10% denatured polyacrylamide sequencing gels (12). PCR was performed with an Expand High Fidelity PCR kit (Roche Diagnostics K.K., Tokyo). The first denaturation step (95°C for 5 min) was followed by 30 cycles, each of 95°C for 30 sec, 60°C for 30 sec, and 72°C for 1 min. Primers were composed of a part of the sequence of the COL1A2 gene (capital letters) and an additional sequence (lower case letters; bold letters indicate recognition sequence of restriction enzymes) for subsequent cloning. H49h1A2S (5′-**gcggtacc**TCATGGGGACCTTAGGC-3′) and H126h1A2A (5′-**gcggtacc**TCTTGGGATGGCATTCC-3′) were used for amplification of the PCR fragment including the upstream repeat. Similarly, H101h1A2S (5′-**gcaagctt**CCACCCACACAGCACGG-3′) and H134h1A2A (5′-**gcgagctc**TAAAAGTGAATGAAGGGGG-3′) were designed for the intron repeat. The upstream repeat could be expressed generally by the formula 5′-$(CA)_l$ $CGCACA(CG)_m(CA)_n$-3′. Here, the alleles observed in this repeat could be expressed as (l, m, n). On the other hand, polymorphism of the intron repeat was restricted to the repetition number of the simple repetitive sequence, GT, so this repeat could be expressed as 5′-(GT)x-3′.

Repeat numbers of some samples including all homozygotes of haplotypes were confirmed by direct sequencing. PCR fragments including either of the dinucleotide repeat regions were amplified, and used as a template for subsequent direct sequencing (12).

Transfection experiments. Cell culture and DNA transfection were performed as described previously (12). In brief, human fibroblasts were co-transfected with firefly luciferase constructs containing COL1A2 promoter region and seapansy luciferase by use of FuGENE 6 reagents (Roche Diagnostics K.K., Tokyo), and relative luciferase activities were expressed as a percentage of constructs showing the lowest activity. Mean ± SD values for eight dishes from two independent experiments were calculated.

Statistical analysis. Haplotypes composed of the combination of upstream and intron repeats were determined by use of the linkage disequilibrium values determined previously (12). The strength of the statistical association between SSc and microsatellite haplotypes in the COL1A2 gene was expressed by relative risk (RR), calculated by the method of Woolf (15). When the case number was null, unity was used to calculate RR. The statistical significance was examined by Chi-square test with Yates' correction. Fisher's exact probability test was used for comparison between groups, when the case number in any given cell was less than 5. Significance was defined as a P value of less than 0.05.

RESULTS

Haplotype Frequency of the COL1A2 Gene

To investigate the possible disease susceptible gene for SSc, we examined the association of the disease with various combinations (repeat-haplotypes) of microsatellites that showed combination-dependent transcriptional stimulation of the COL1A2 gene (12). When we compared the frequencies of repeat-haplotypes, that is, specific combinations of upstream and intron repeats, we found 3 homozygotes of the (13,6,8)-12 haplotype among 93 patients but none in 209 controls. The relative frequency of the homozygote in the patients was significantly higher than that in the healthy controls, 0.032 versus 0.00, $P = 0.029$, RR > 6.93 (Table 1). The statistical significance of the association between the homozygote of this haplotype with the disease was increased, when the frequency of this haplotype in the patient group positive for ANAs, anti-DNA topo-isomerase I antibody, anti-centromere antibody, or anti-U1-ribonucleoprotein antibody, was compared with that for the controls, from $P = 0.029$ to $P = 0.014$. The highest association between the homozygote and SSc was observed with male patients positive for the ANAs, the relative frequency was 0.133 versus 0.00, $P = 0.0042$, RR > 32.0 (Table 1). We also analyzed a similar number (91 cases) of non-SSc patients, i.e., patients with dilated cardiomyopathy (Akai *et al.*, unpublished data); but again there was no homozygote of the haplotype. Thus P value was further

TABLE 1

Association of Homozygous Repeat-Haplotypes (Functional Microsatellites) in the Human COL1A2 Gene with Systemic Sclerosis (SSc)

Haplotype (l,m,n)-x	Healthy controls (N = 209*[1])	SSc (N = 93)	SSc (TCR+*[2]) (N = 68)	SSc (male) (N = 23)	SSc (male) (TCR+) (N = 15)
Total	36 (0.172*[3])	12 (0.129)	10 (0.147)	5 (0.261)	4 (0.266)
P*[4]		0.437	0.766	0.380	0.189
RR*[5]		0.711	0.829	1.330	2.130
(14,6,8)-17					
Frequency	25 (0.120)	6 (0.065)	4 (0.059)	3 (0.130)	2 (0.133)
P		0.206	0.111	0.549	0.565
RR		0.505	0.458	1.098	1.126
(15,7,8)-12					
Frequency	11 (0.052)	3 (0.032)	3 (0.044)	0 (0)	0 (0)
P		0.326	0.536	0.308	0.458
RR		0.600	0.831	0	0
(14,6,8)-17+ (15,7,8)-12					
Frequency	36 (0.172)	9 (0.097)	7 (0.103)	3 (0.130)	2 (0.133)
P		0.127	0.238	0.435	0.515
RR		0.510	0.551	0.721	0.739
(13,6,8)-12					
Frequency	0 (0)	3 (0.032)	3 (0.044)	2 (0.087)	2 (0.133)
P		0.029#	0.014#	0.0095#	0.0042#
RR		>6.93	>9.60	>19.81	>32.00
Upstream repeat (l,m,n) = (13,6,8)					
Frequency	2 (0.010)	4 (0.043)	4 (0.058)	2 (0.087)	2 (0.133)
P		0.075	0.034#	0.050	0.023#
RR		4.63	6.44	9.81	15.85
Intron repeat (x) = 12					
Frequency	51 (0.245)	26 (0.280)	22 (0.323)	7 (0.304)	5 (0.333)
P		0.625	0.266	0.713	0.651
RR		1.19	1.47	1.35	1.54

*[1] Number of subjects examined in each group.
*[2] SSc patients positive for anti-DNA topoisomerase I antibody, anti-centromere antibody, or anti-U1 ribonucleoprotein antibody.
*[3] Frequency relative to total number of subjects.
*[4] P values were calculated by Chi-square analysis with Yates' correction, and Fisher's exact probability test was employed when the case number in any given cell was less than 5.
*[5] Relative risk (RR) was calculated by the method of Woolf. Unity was tentatively used when the case number in any given cell was null.
Significantly different from controls.

decreased and RR for the disease was increased when the values were compared between male SSc patients positive for the ANAs and non-SSc subjects.

This association indicates that the COL1A2 gene itself is a disease gene and not reflection of a simple linkage-disequilibrium between the microsatellite and the disease because we could only find lower association (upstream repeats) or no association (intron repeats) of either of the repeats, which constitute the haplotype, with the disease (Table 1).

Transcriptional Activity of the Constructs Containing COL1A2 Gene Fragments

The rates of transcriptional stimulation of the COL1A2 gene were different depending on the combinations of the repeats or repeat-haplotypes (12). Thus we compared the rates of stimulation and the relative frequency of homozygotes for other repeat-haplotypes found in patients positive for the ANAs to those in controls.

The frequency of the homozygous individuals having the haplotype (13,6,8)-12 was significantly higher in male SSc patients positive for the ANAs than in the controls, and these individuals had a high risk factor (Fig. 1). A smaller portion of the ANAs positive male patients had haplotypes showing the lower transcriptional stimulation activities indicated a lower relative frequency (0.133) than the controls (0.172), though the difference was not significant (P = 0.53, Fig. 1). Thus the RR of the homozygous subjects further increased for SSc than that of the subjects containing the haplo-

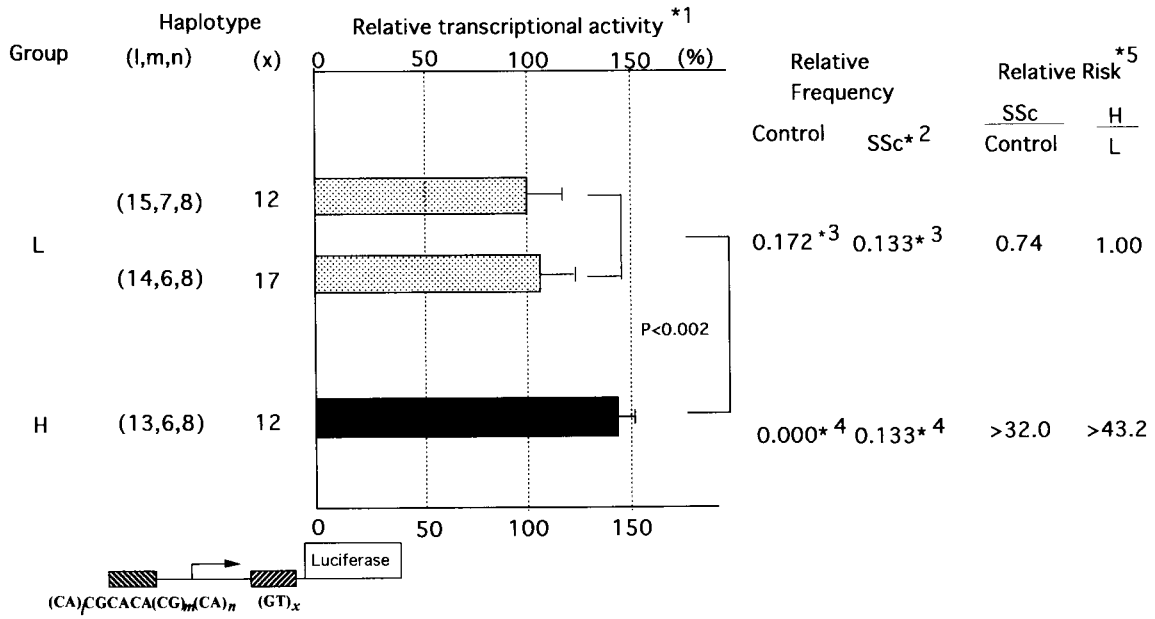

FIG. 1. Relationship between frequency of homozygous haplotypes and their transcription-stimulating activities. (*1) Transcriptional activity was measured by luciferase activity after transfection of human fibroblasts with the constructs containing various haplotypes. (*2) Male patients positive for the ANAs. (*3) Subjects containing haplotypes with low transcription-stimulating activities. $P = 0.53$ (*4) Subjects homozygous for the haplotype (13,6,8)-12. $P = 0.0042$ (*5) Relative risk was calculated by the method of Woolf. Unity was tentatively used when the case number in any given cell was null.

types with the lower transcriptional stimulation rates, from >32.0 to >43.2 (Fig. 1).

DISCUSSION

Recently, it was shown that microsatellites are not merely markers for genes but also have some physiological functions such as regulation of gene expression (16). This is true for the human COL1A2 gene. We recently found that the co-presence of two dinucleotide repeats near the transcriptional start site of the gene, but not either alone, enhanced transcriptional activity of COL1A2 gene (12). Here we investigated the possible association of these repeats with SSc and found that homozygotes of a specific combination of the repeats, (13,6,8)-12 were predominantly found in the male patients expressing ANAs.

Higher association of the homozygotes with the disease than the heterozygotes suggests that the effect is phenotypically recessive. The increase in significance of the association with the disease those among patients positive for ANAs suggests that the induction of immunological disturbance such as production of autoantibodies may be closely related to the development of the disease in association with the COL1A2 haplotype. It is reported that the expression of ANAs such as anti-topo I antibody is partly predetermined by HLA haplotypes (2, 3), suggesting susceptibility to the disease is determined by a combination of different genes, in this case COL1A2 gene and HLA genes.

In a population of full-blooded Choctaw Native Americans, who are relatively genetically homogeneous, a high prevalence of SSc was found only in the residents of southeastern Oklahoma counties (8), suggesting environmental factors might be also responsible for the expression of the disease. These would be reasons why we could not find other cases of SSc in the families of the homozygotes and rarely find familial SSc. The increased association of the disease in male patients may depend on factors such as the hormonal imbalance and/or graft versus host reaction-like phenomenon occurs more often in female patients (17).

This paper is the first report that indicates the significant association of a specific combination of functional dinucleotide repeats in the human COL1A2 gene with SSc.

The cases reported here account for only three (3/93) percent of the patients investigated, indicating a genetic heterogeneity and a multi-factorial basis for the disease, but this study revealed a new aspect for understanding this complex disease.

ACKNOWLEDGMENTS

We thank Dr. Mieko Yanokura and Ms. Kazuyo Shirai (Laboratory of Biological Information) for DNA sequencing. This work was supported in part by the following: a grant from the Intractable Disease Division (Scleroderma Research Committee), Public Health Bureau, the Ministry of Health and Welfare of Japan, and research grants from the Nakatomi Foundation and Terumo Life Science Foundation.

REFERENCES

1. Medsger, T. A., Jr. (1997) *in* Arthritis and Allied Conditions (Koopman, W. J., Ed.), 13th ed., pp. 1433–1464, Williams and Wilkins, Philadelphia, PA.
2. Kuwana, M., Kaburaki, J., Okano, Y., Inoko, H., and Tsuji, K. (1995) *J. Clin. Invest.* **92,** 1296–1301.
3. Arnett, F. C. (1995) *Intern. Rev. Immunol.* **12,** 107–128.
4. Jimenez, S. A., and Christner, P. (1994) *Clin. Dermatol.* **12,** 425–436.
5. Siracusa, L. D., McGrath, R., Ma, Q., Moskow, J. J., Manne, J., Christner, P. J., Buchberg, A. M., and Jimenez, S. A. (1996) *Genome Res.* **6,** 300–313.
6. Sheldon, W. B., Lurie, D. P., Maricq, H. R., Kahaleh, M. B., DeLustro, F. A., Gibofsky, A., and LeRoy, E. C. (1981) *Arthritis Rheum.* **24,** 668–676.
7. McGregor, A. R., Watson, A., Yunis, E., Pandey, J. P., Takehara, K., Tidwell, J. T., Ruggieri, A., Silver, R. M., LeRoy, E. C., and Maricq, H. R. (1988) *Am. J. Med.* **84,** 1023–1032.
8. Tan, F. K., Stivers, D. N., Foster, M. W., Chakraborty, R., Howard, R. F., Milewicz, D. M., and Arnett, F. C. (1998) *Arthritis Rheum.* **41,** 1729–1737.
9. Bauer, E. A., Cruz, D. J. S., Uitto, J., and Eisen, A. Z. (1987) *in* Connective Tissue Disease (Uitto, J., and Perejda, A. J., Eds.), pp. 249–261, Dekker, New York.
10. Hata, R. I., Horikawa, S., Kurata, S. I., Senoo, H., and Tsukada, K. (1994) *in* Pathogenesis and Management of Scleroderma and Connective Tissue Disorders (Nishioka, K., and Krieg, T., Eds.), pp. 5–13, Scleroderma Research Committee, Tokyo.
11. Akai, J., Kimura, A., Arai, K., Uehara, K., and Hata, R. I. (1998) *Connect. Tiss.* **30,** 1–6.
12. Akai, J., Kimura, A., and Hata, R. I. (1999) *Gene* **239,** 65–73.
13. Subcommittee for Scleroderma Criteria of the American Rheumatism Association Diagnostic and Therapeutic Criteria Committee. Preliminary criteria for the classification of systemic sclerosis (scleroderma) (1980) *Arthritis Rheum.* **23,** 581–590.
14. Kuwana, M., Kaburaki, J., Medsger, T. A., Jr., and Wright, T. M. (1999) *Arthritis Rheum.* **42,** 1179–1188.
15. Woolf, B. (1955) *Ann. Hum. Genet.* **19,** 251–253.
16. Kashi, Y., King, D., and Soller, M. (1997) *Trends Genet.* **13,** 74–78.
17. Artlett, C. M., Welsh, K. I., Black, C. M., and Jimenez, S. A. (1997) *Immunogenetics* **47,** 17–22.