SHORT COMMUNICATION

Conserved Worldwide Linkage Disequilibrium in the Human Factor XI Gene

Takashi Tarumi,*,†,[1] Danko Martincic,‡,[1] James A. Whitlock,‡ Jonathan H. Addy,§ Scott M. Williams,¶ and David Gailani*,†,[2]

*Department of Pathology, †Department of Medicine, and ‡Department of Pediatrics, Vanderbilt University, Nashville Tennessee 37232-6305; §Department of Medicine and Therapeutics, University of Ghana Medical School, Accra, Ghana; and ¶Department of Microbiology, Meharry Medical College, Nashville, Tennessee 37208

Received July 20, 2000; accepted September 12, 2000; published online November 13, 2000

We have identified, in four diverse human populations, five common single-nucleotide polymorphisms (SNPs) in the coding region of the gene for the blood coagulation protease factor XI. Each SNP has an allele frequency >5% in at least one population. Three of the SNPs (C472T, A844G, and T1234C), spread out over approximately 10 kb of genomic DNA, are in marked linkage disequilibrium (LD) with one another ($P < 10^{-4}$). Interestingly, haplotypes associated with the linked SNPs are conserved across all populations studied, despite significantly different allele frequencies between populations. The presence of such common, widely dispersed haplotypes could complicate the interpretation of LD studies and emphasizes the need for a better understanding of general patterns of LD to facilitate identification of genes for common disorders.

Single-nucleotide polymorphisms (SNPs) are the most common type of sequence variation in the human genome. While identification of SNPs has shed light on many Mendelian diseases, they are also powerful tools for identifying genomic regions associated with common complex disorders and for explaining evolutionary history (4, 5, 13, 15, 16). Techniques such as linkage disequilibrium (LD) analysis can demonstrate correlations between diseases and SNPs, facilitating discovery of disease-associated genes. To this end, a major effort is under way to catalog the genetic variation in humans and to correlate variation with pathology (4, 13, 15). This work is based on the premise that the genetic milieu of a disease-predisposing mutation does not change appreciably over short periods of time nor over short distances along a chromosome. While this is undoubtedly true, LD analysis is hampered by a lack of data concerning the extent of LD as a function of haplotype age, distance along a chromosome, and regional variation in recombination rates (3, 7, 9). Indeed, there is evidence that recombination rates vary considerably even across small regions (8, 12). Selection pressure or small population sizes also influence LD (23). In the absence of experimental data for multiple genes and populations, models of LD make critical assumptions for factors that may vary widely across the genome (6, 15, 20).

The gene for the human blood coagulation protease factor XI (F11) contains 15 exons spanning ~25 kb on the long arm of chromosome 4 (4q35) (1, 14). Recently, we identified three SNPs in the coding region of this gene (C472T, A844G, and T1234C; Fig. 1A) and demonstrated that the minor alleles are common (18). We have extended this study to 173 normal individuals from diverse human populations and show that these SNPs are in strong LD with one another. Furthermore, haplotype patterns involving the SNPs are maintained across populations, despite significant allele frequency differences. Coding and flanking intronic regions of 266 F11 alleles from 41 Caucasians, 50 West Africans (Ghanaians), and 42 East Asians (Japanese, Chinese, and Koreans) were amplified by PCR and screened for sequence variation by dideoxyfingerprinting (18). Direct sequencing of PCR fragments then definitively identified the SNPs. Eleven SNPs were identified in these 133 individuals; four of the SNPs are in noncoding regions. Four SNPs were present in only one individual, one was present in two individuals, and another was present in five East Asian individuals. The remaining five SNPs, all of which are in exons (hereafter referred to as cSNPs; Fig. 1A), had minor allele frequencies >5% in at least one ethnic population (Fig.

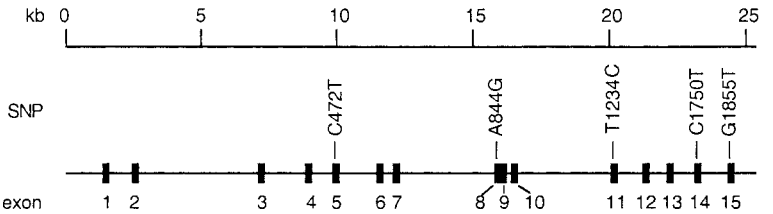
The Institutional Review Boards of Vanderbilt University and Meharry Medical College approved procedures for obtaining blood samples and processing DNA from healthy volunteers.

[1] These authors contributed equally to this work.

[2] To whom correspondence should be addressed at Division of Hematology/Oncology, Vanderbilt University, 538 MRB II, 2220 Pierce Avenue, Nashville, TN 37232-6305. Telephone: (615) 936-1505. Fax: (615) 936-3853. E-mail: dave.gailani@mcmail.vanderbilt.edu.

Genomics **70,** 269–272 (2000)
doi:10.1006/geno.2000.6393
0888-7543/00 $35.00

A



B

	SNP[1]				
	C472T[2]	A844G[2]	T1234C[3]	C1750T[4]	G1855T[5]
Caucasian (n=82)	0.05 (4)	0.13 (11)	0.13 (11)	0.01 (1)	0.06 (5)
Ghanaian (n=100)	0.19 (19)	0.25 (25)	0.24 (24)	0.16 (16)	0.09 (9)
Oriental (n=84)	0.00 (0)	0.06 (5)	0.06 (5)	0.02 (2)	0.06 (5)
Ethiopian (n=80)	0.11 (9)	0.31 (25)	0.26 (21)	ND	ND
Total (n=346)	0.09 (32)	0.18 (66)	0.17 (61)	ND	ND

[1] Allele frequency differences were calculated using the algorithm of Raymond and Rousset (22) and Tools for Population Genetic analysis software version 1.3 (19).
[2] $p < 10^{-4}$
[3] $p = 0.0026$
[4] $p = 0.0001$
[5] $p = 0.0559$
ND - not determined

FIG. 1. (**A**) Schematic diagram of the human factor XI (F11) gene, demonstrating the relative positions of, and absolute distances between, exons (1). The position of five common SNPs are shown above the exons, with the nucleotide representing the predominant allele preceding the basepair position number and the nucleotide representing the minor allele following the basepair number. (**B**) Allele frequencies of the SNPs in the coding sequence of the F11 gene. The numbering system used in both parts of the figure refers to the position in the published sequence of the human F11 cDNA (10).

1B). They are the subjects of this report. None of the cSNPs encodes an amino acid change. Three cSNPs are identical to those identified in our earlier study (C472T, A844G, and T1234C) (18), and all five were reported by Cargill *et al.* (4). For all cSNPs, the sequence of the common allele is identical to the published cDNA sequence (10) and to the chimpanzee gene sequence (http://www.genome.wi.mit/edu/cvar_snps) (4), indicating that they represent the ancestral allele. The cSNPs individually have genotype frequencies that are in Hardy–Weinberg equilibrium (analysis not shown), and all are present in the three ethnic populations (with the exception of 472T in East Asians), indicating origins predating the exit of modern *Homo sapiens* from Africa to Asia and Europe.

Interestingly, the C472T, A844G, and T1234C cSNPs, which are spread across ~10 kb of DNA (Fig. 1A), are in significant LD with one another (Tables 1A and 1B). Tests for LD were performed using GDA software (17) based on the methods of Weir (25). In these populations, all pairwise and three-way comparisons between C472T, A844G, and T1234C (except for East

TABLE 1A

Genotype Frequencies in the Human Factor XI Gene by Ethnic Group for cSNPs C472T, A844G, and T1234C

SNP						
472	844	1234	Caucasian (n = 41)	Ghanaian (n = 50)	Oriental (n = 42)	Ethiopian (n = 40)	Total (n = 173)
C/C	A/A	T/T	0.76 (31)	0.58 (29)	0.88 (37)	0.45 (18)	0.66 (115)
C/T	A/A	T/T	0.00 (0)	0.00 (0)	0.00 (0)	0.00 (0)	0.00 (0)
C/C	A/G	T/T	0.00 (0)	0.02 (1)	0.00 (0)	0.10 (4)	0.03 (5)
C/C	A/A	T/C	0.00 (0)	0.00 (0)	0.00 (0)	0.03 (1)	0.01 (1)
C/T	A/G	T/T	0.00 (0)	0.00 (0)	0.00 (0)	0.00 (0)	0.00 (0)
C/C	A/G	T/C	0.14 (6)	0.08 (4)	0.12 (5)	0.18 (7)	0.13 (22)
C/T	A/A	T/C	0.00 (0)	0.00 (0)	0.00 (0)	0.00 (0)	0.00 (0)
C/T	A/G	T/C	0.10 (4)	0.24 (12)	0.00 (0)	0.13 (5)	0.12 (21)
C/T	G/G	T/C	0.00 (0)	0.00 (0)	0.00 (0)	0.05 (2)	0.01 (2)
C/C	A/G	C/C	0.00 (0)	0.00 (0)	0.00 (0)	0.03 (1)	0.01 (1)
C/T	G/G	C/C	0.00 (0)	0.02 (1)	0.00 (0)	0.05 (2)	0.02 (3)
T/T	G/G	C/C	0.00 (0)	0.06 (3)	0.00 (0)	0.00 (0)	0.02 (3)

TABLE 1B

A Comparison of Estimated Haplotype Frequencies and Random Expectation for cSNPs C472T, A844G, and T1234C in the Human Factor XI Gene

SNPs			Haplotype frequencies									
			Caucasian		Ghanaian		Oriental		Ethiopian		Total	
472	844	1234	Estimated[a]	Expected	Estimated[a]	Expected	Estimated[a]	Expected	Estimated[a]	Expected	Estimated[a]	Expected
C	A	T	0.865	0.713	0.750	0.462	0.940	0.884	0.660	0.449	0.803	0.605
T	A	T	0.000	0.036	0.000	0.108	0.000	0.000	0.000	0.057	0.000	0.062
C	G	T	0.000	0.110	0.010	0.153	0.000	0.055	0.077	0.204	0.020	0.142
C	A	C	0.000	0.110	0.000	0.145	0.000	0.055	0.027	0.160	0.006	0.129
T	G	T	0.000	0.005	0.000	0.036	0.000	0.000	0.000	0.025	0.000	0.014
T	A	C	0.000	0.056	0.000	0.034	0.000	0.000	0.000	0.020	0.000	0.013
C	G	C	0.085	0.017	0.050	0.048	0.059	0.003	0.122	0.072	0.078	0.030
T	G	C	0.048	0.001	0.190	0.011	0.000	0.000	0.112	0.009	0.092	0.003

Note. There are eight possible haplotypes for the three sites.

[a] Estimated haplotype frequencies are highly significantly different from expected for all groups, with P values $<10^{-4}$. Estimated haplotype frequencies were generated with the EH program (24). The expected frequencies were generated assuming linkage equilibrium.

Asians in whom the 472 T allele was absent) showed significant deviation from equilibrium at least at the $P = 0.005$ level and as low as $P = 10^{-6}$ (analysis not shown). In comparison, cSNPs C1750T and G1855T do not demonstrate LD relative to other cSNPs in any population (analysis not shown) or to each other despite being separated by <2 kb (Fig. 1A). In the three populations, 844G and 1234C are always found together, with the exception of one Ghanaian, while 472T is always associated with 844G–1234C. Estimated haplotype frequencies for the three populations were determined using EH software (24, 27). This algorithm estimates haplotypes based on genotype data and compares them to haplotype frequencies generated under the assumption that there is no association between sites. The haplotype frequencies (with and without association) are compared using a χ^2 distribution. Haplotype frequencies for all three populations are significantly different from random expectations ($P < 10^{-4}$; Table 1B). We analyzed DNA from an additional 40 individuals of East African (Ethiopian Jewish) ancestry. SNPs 472T, 844G, and 1234C are present at frequencies similar to those in Ghanaians (Fig. 1B) and again are in significant LD relative to one another ($P < 10^{-4}$) (Tables 1A and 1B). The Ethiopian haplotype patterns are similar to those from other populations. While 844G or 1234C is found in isolation in a few Ethiopians, the 844G–1234C combination predominates, and 472T is always associated with 844G–1234C.

We were interested in determining whether common mutations causing congenital deficiency of F11 were associated with particular haplotypes. F11 deficiency is an autosomal disorder characterized by excessive bleeding after trauma or surgery and is prevalent in persons of Jewish ancestry (2, 21). Over 90% of abnormal F11 alleles in this population have one of two mutations, designated types II (Glu117Ter) and III (Phe283Leu). Peretz *et al.* demonstrated that each mutation is associated with a single haplotype and therefore represents a single mutation event (21). We analyzed DNA from a homozygote for the type II mutation and determined that this person is also homozygous for the 472T–844G–1234C haplotype. This demonstrates that 472T, 844G, and 1234C are on the same chromosome in this individual. Portions of F11 alleles from a heterozygote for the type II mutation, and a compound heterozygote for the type II and III mutations, were subcloned and analyzed for cSNPs. The analysis confirmed that the type II mutation and the 472T–844G–1234C haplotype are on the same chromosome, while the type III mutation is on the chromosome with the common C472–A844–T1234 haplotype.

Given the apparent old age of the 472T–844G–1234C haplotype, and the distances between the loci (~10 kb), some models predict that LD between the sites should have dissipated (15). Two other cSNPs (C1750T and G1855T) are common in the populations we studied, and are in equilibrium with each other (despite being separated by <2 kb) and with the 472T–844G–1234C haplotype. This suggests that sufficient time has passed for initial LD to decay in this region in the absence of a process limiting recombination. The cause of LD in this case is not clear. It could be due to inherent variability in recombination rates across the finite length of the F11 gene. Cullen *et al.* (8) demonstrated a wide range of recombination rates across the HLA class II region, with the highest rates clustered in three relatively small areas (8.8–50 kb) of the 500-kb interval studied. Alternatively the data are consistent with a strong selection pressure on the 472T–844G–1234C haplotype. No phenotype has been linked to this haplotype, and it is not associated with amino acid changes. We are investigating the possibility that the haplotype influences plasma levels of F11 by affecting F11 mRNA translation or stability. A peculiar feature of the 472T–844G–1234C haplotype is its striking conservation across populations, despite significantly different allele frequencies between populations. This in-

dicates that selection, if present, is acting on the whole haplotype. This is in stark contrast to the majority of published data showing that specific haplotypes are not usually found across a wide array of populations (7, 11). For example, the single basepair change in the β-globin gene that causes sickle cell anemia is under strong selection pressure because of its protective effect against malaria, but is found on several different haplotype backgrounds in different populations (26).

Our data demonstrate marked LD in the human F11 gene. Just as striking is the observation that the haplotypes involved are conserved across diverse populations. Regardless of the cause of LD in this case, the ubiquity of a haplotype such as F11 472T–844G–1234C has major implications for using LD analysis to identify disease-associated genes. For example, both the F11 type II and III mutations cause similar clinical syndromes but are associated with different common F11 haplotypes. The presence of two common disease-associated haplotypes would make it difficult to localize a genetic abnormality in this disorder by LD analysis in the absence of other compelling data. This finding supports the suggestion of Clark *et al.* (7) that "the design and interpretation of disease association studies may not be as straightforward as is often assumed" and underscores the need for additional information about the genetic structure of human populations.

ACKNOWLEDGMENTS

The authors thank Drs. K. Morgan, J. Moore, J. Phillips, S. Russell, and S. Tishkoff for helpful discussions during preparation of the manuscript. This work was supported by grants R01 HL58837 and K14-HL03321 from the National Heart, Lung, and Blood Institute; by Grant G12-RR03032 from the National Center for Research Resources; by Grant T37-TW00043 Fogarty Center Grant; and by the Andrew Mellon Foundation. David Gailani is an Established Investigator of the American Heart Association.

REFERENCES

1. Asakai, R., Davie, E., and Chung, D. (1987). Organization of the gene for human factor XI. *Biochemistry* **26:** 7221–7228.
2. Asakai, R., Chung, D., Davie, E., and Seligsohn, U. (1991). Factor XI deficiency in Ashkenazi Jews in Israel. *N. Eng. J. Med.* **325:** 153–158.
3. Broman, K., and Weber, J. (1999). Long homozygous chromosomal segments in the CEPH families. *Am. J. Hum. Genet.* **65:** 1493–1500.
4. Cargill, M., Altshuler, D., Ireland, J., Sklar, P., Ardlie, K., Patil, N., Lane, C. R., Lim, E. P., Kalyanaraman, N., Nemesh, J., Ziaugra, L., Friedland, L., Rolfe, A., Warrington, J., Lipshutz, R., Daley, G. Q., and Lander, E. S. (1999). Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nat. Genet.* **22:** 231–238.
5. Chakravarti, A. (1999). Population genetics—Making sense out of sequence. *Nat. Genet.* **21:** 56–60.
6. Clark, A. (1999). The size distribution of homozygous segments in the human genome. *Am. J. Hum. Genet.* **65:** 1489–1492.
7. Clark, A., Weiss, K., Nickerson,D., Taylor, S., Buchanan, A., Stengard, J., Salomaa, V., Vartiainen, E., Perola, M., Boerwinkle, E., and Sing, C. F. (1998). Haplotype structure and population genetic inferences from nucleotide-sequence variation in human lipoprotein lipase. *Am. J. Hum. Genet.* **63:** 595–612.
8. Cullen, M., Nobel, J., Erlich, H., Thorpe, K., Beck, S., Klitz, W., Trowsdale, J., and Carrington, M. (1997). Characterization of recombination in the HLA class II region. *Am. J. Hum. Genet.* **60:** 397–407.
9. Freimer, N. B., Service, S. K., and Slatkin, M. (1997). Expanding on population studies. *Nat. Genet.* **17:** 371–373.
10. Fujikawa, K., Chung, D., Hendrickson, L., and Davie, E. (1986). Amino acid sequence of human factor XI, a blood coagulation factor with four tandem repeats that are highly homologous with plasma prekallikrein. *Biochemistry* **25:** 2417–2424.
11. Goddard, K., Hopkins, P., Hall, J., and Witte, J. (2000). Linkage disequilibrium and allele-frequency distributions for 114 single-nucleotide polymorphisms in five populations. *Am. J. Hum. Genet.* **66:** 216–234.
12. Gordon, D., Simonic, I., and Ott, J. (2000). Significant evidence for linkage disequilibrium over a 5-cM region among Afrikaners. *Genomics* **66:** 87–92, doi:10.1006/geno.2000.6190.
13. Halushka, M., Fan, J-B., Bentley, K., Hsie, L., Shen, N., Weder, A., Cooper, R., Lipshutz, R., and Chakravarti, A. (1999). Patterns of single-nucleotide polymorphisms in candidate genes for blood-pressure homeostasis. *Nat. Genet.* **22:** 239–247.
14. Kato, A., Asakai, R., Davie, E., and Aoki, N. (1989). Factor XI gene (F11) is located on the distal end of the long arm of human chromosome 4. *Cytogenet. Cell. Genet.* **52:** 77.
15. Kruglyak, L. (1999). Prospects for whole-genome linkage disequilibrium mapping of common disease genes. *Nat. Genet.* **22:** 139–144.
16. Laan, M., and Pääbo, S. (1997). Demographic history and linkage disequilibrium in human populations. *Nat. Genet.* **17:** 435–438.
17. Lewis, P. O., and Zaykin, D. (2000). Genetic Data Analysis: Computer program for the analysis of allelic data. Version 1.0 (d12) (http://alleyn.eeb.uconn.edu/gda/).
18. Martincic, D., Zimmerman, S., Ware, R., Sun, M., Whitlock, J., and Gailani, D. (1998). Identification of mutations and polymorphisms in the factor XI genes of an African-American family by dideoxyfingerprinting. *Blood* **92:** 3309–3317.
19. Miller, M. P. (1997). Tools for population genetic analysis (TFPGA) 1.3: A windows program for the analysis of allozyme and molecular population genetic data. Computer software distributed by author (http://herb.bio.nau.edu/~miller/tfpga.htm).
20. Ott, J. (2000). Predicting the range of linkage disequilibrium. *Proc. Natl. Acad. Sci. USA* **97:** 2–3.
21. Peretz, H., Mulai, A., Usher, S., Zivelin, A., Segal, A., Weisman, Z., Mittleman, M., Lupo, H., Lanir, N., Brenner, B., Shpilberg, O., and Seligsohn, U. (1997). The two common mutations causing factor XI deficiency in Jews stem from distinct founders: One of ancient Middle Eastern origin and another of more recent European origin. *Blood* **90:** 2654–2659.
22. Raymond, M. L., and Rousset, F. (1995). An exact test for population differentiation. *Evolution* **49:** 1280–1283.
23. Slatkin, M. (2000). Balancing selection at closely linked, overdominant loci in a finite population. *Genetics* **154:** 1367–1378.
24. Terwilliger, J., and Ott, J. (1994). "Handbook of Human Genetics," Johns Hopkins Univ. Press, Baltimore.
25. Weir, B. (1996). "Genetic Data Analysis II," Sinauer Associates, Sunderland, MA.
26. Weiss, K. M. (1993). "Genetic Variation and Human Disease: Principles and Evolutionary Approaches," Cambridge Univ. Press, New York.
27. Xie, X., and Ott, J. (1993). Testing linkage disequilibrium between a disease gene and marker loci. *Am. J. Hum. Genet.* **53:** 1107.