Am. J. Hum. Genet. 64:1071–1075, 1999

Age Estimates of Two Common Mutations Causing Factor XI Deficiency: Recent Genetic Drift Is Not Necessary for Elevated Disease Incidence among Ashkenazi Jews

David B. Goldstein,[1] David E. Reich,[2] Neil Bradman,[2] Sali Usher,[3] Uri Seligsohn,[4] and Hava Peretz[3,4]

[1]Galton Laboratory, Department of Biology, University College London, London; [2]Department of Zoology, University of Oxford, Oxford; [3]Department of Clinical Biochemistry, Sourasky Medical Center, Tel Aviv; and [4]Department of Haematology, Institute of Thrombosis and Hemostasis, Chaim Sheba Medical Center, Tel Hashomer, Israel

Summary

The type II and type III mutations at the *FXI* locus, which cause coagulation factor XI deficiency, have high frequencies in Jewish populations. The type III mutation is largely restricted to Ashkenazi Jews, but the type II mutation is observed at high frequency in both Ashkenazi and Iraqi Jews, suggesting the possibility that the mutation appeared before the separation of these communities. Here we report estimates of the ages of the type II and type III mutations, based on the observed distribution of allelic variants at a flanking microsatellite marker (*D4S171*). The results are consistent with a recent origin for the type III mutation but suggest that the type II mutation appeared >120 generations ago. This finding demonstrates that the high frequency of the type II mutation among Jews is independent of the demographic upheavals among Ashkenazi Jews in the 16th and 17th centuries.

Introduction

Coagulation factor XI deficiency results from mutations at the *FXI* gene located on chromosome 4. The disease is common in Ashkenazi Jews, present to a lesser degree among other Jews, and occurs only rarely in non-Jewish populations (Seligsohn 1978; Saito et al. 1985; Bolton-Maggs et al. 1992). Four distinct mutations are observed in Jewish populations, with types II and III by far the most frequent. The type III mutation, caused by a missense change at position 283, has an allele frequency of 2.54% in Ashkenazi Jews but has not been observed in large samples of Iraqi and Sephardic Jews (Shpilberg et al. 1995; Peretz et al. 1997). The type II mutation, caused by a nonsense change at position 117, has a frequency of 2.17% in Ashkenazi Jews, occurs with a frequency of 1.67% in Iraqi Jews, and is present at lower frequencies in other Jewish populations. A common founder for this mutation has been indicated by the presence of a single background haplotype among both Ashkenazi and Iraqi Jews (Peretz et al. 1997). Of the many single-locus disorders with high frequency among Jews, the type II factor XI mutation is the first to be reported with high frequency in both Ashkenazi and non-Ashkenazi Jewish populations. The origin of Iraqi Jewry is controversial, but it has been argued that it derives from the Babylonian exile, >2,500 years ago. According to this view, the type II mutation may have occurred within the ancestral Jewish population, before the Babylonian exile and well before the dispersion after the Bar Kochba revolt against the Romans (132–135 A.D.). The type III mutation, being restricted to Ashkenazi Jews, is suggested to have arisen sometime after the emergence of Ashkenazi Jews and in substantial isolation from Iraqi Jews (Shpilberg et al. 1995). To test these predictions, we used the observed distribution of allelic variants of a flanking microsatellite marker (*D4S171*) on mutant and control backgrounds to estimate the age of each mutation (Reich and Goldstein, in press; Stephens et al. 1998).

Data and Analysis

When a new mutation appears, it occurs on a single chromosome and thereby generates linkage disequilibrium with any polymorphic markers. The rate at which this disequilibrium breaks down depends on the recombination and mutation rates; hence, if estimates of these

Received July 16, 1998; accepted for publication January 22, 1999; electronically published March 15, 1999.

Address for correspondence and reprints: Dr. David B. Goldstein, Galton Laboratory, Department of Biology, University College London, Wolfson House, 4 Stephenson Way, London NW1 2HE, United Kingdom. E-mail: d.goldstein@ucl.ac.uk; or Dr. Hava Peretz, Department of Clinical Biochemistry, Sourasky Medical Center, Tel Aviv 64239, Israel.

0002-9297/99/6404-0019$02.00

are available, the observed level of disequilibrium can be used to estimate the date of the mutation (Risch et al. 1995; Stephens et al. 1998). To achieve statistical power, however, it is important to use a marker at an appropriate recombinational distance—that is, great enough to ensure the generation of nonancestral haplotypes at moderate frequency but, ideally, not so great as to equilibrate allele frequencies within the time frame of interest. Of 74 informative meioses, we observed a single recombination event between *FXI* and *D4S171* (a dinucleotide microsatellite), implying a recombination rate of .0135, appropriate for estimating coalescent times on the order of hundreds of generations. Frequencies of alleles at this marker were determined in 99 chromosomes carrying the type II mutation and in 73 chromosomes carrying the type III mutation. We also characterized 103 chromosomes from healthy Ashkenazi Jews, to estimate the proportion of recombination events that are expected to result in a nonancestral haplotype (table 1).

Table 1

Frequencies of Alleles at the Marker Locus *D4S171* in Affected and Control Chromosomes

Allele Size (bp)	Frequency			
	CEPH[a]	Control (n = 103)	Type II (n = 99)	Type III (n = 73)
143	.01	.02	.00	.00
145	.01	.03	.03	.26
147	.09	.06	.02	.05
149	.04	.08	.02	.01
151	.34	.38	.48	.52
153	.39	.22	.27	.10
155	.08	.12	.06	.01
157	.02	.00	.02	.01
159	.00	.05	.07	.01
161	.01	.00	.01	.00
163	.00	.02	.01	.00
165	.00	.03	.00	.01

[a] Data provided by James Weber.

In their study of idiopathic torsion dystonia, Risch et al. (1995) used a method for estimating the coalescent time (G) of mutant chromosomes that focuses on the proportion of lineages not having undergone a mutation or recombination event. If no regeneration of ancestral haplotypes is assumed, the expected proportion of ancestral haplotypes is effectively $p = e^{-Gr}$, where G is the number of generations since the coalescence of the sampled chromosomes and r is the effective mutation and recombination rate (see eq. [1]). When there is a moderate to high proportion of nonancestral alleles among the mutant chromosomes, however, it is necessary to model the regeneration of ancestral haplotypes by the recombination process. This can be done in a number of ways, but we favor the Markov model representation of Reich and Goldstein (in press) for its flexibility. Under this formulation, it is straightforward to include an arbitrary number of markers and both recombination and mutation.

The full behavior of the system is easily represented as a Markov process in which the state space is the proportion of chromosomes in the ancestral and nonancestral categories, and the transition matrix **K** gives the probabilities that each haplotype will be transformed into the other in a single generation (Stephens et al. 1998; Reich and Goldstein, in press). The transition matrix **K** is given by:

$$\mathbf{K} = c\mathbf{R} + u\mathbf{M} + (1 - c - u)\mathbf{I}\ , \qquad (1)$$

where c and u are scalars reflecting the recombination and mutation rates, respectively. In this case, u = .00056 (Weber and Wong 1993) and c = .0135. The matrices **R** and **M** reflect the probabilities of producing nonancestral haplotypes should a recombination or mutation event occur, whereas **I** is the identity matrix. If a represents the frequency of the ancestral allele in the control population, the matrix **R** has the elements $R_{11} = a$, $R_{12} = a$, $R_{21} = 1 - a$, and $R_{22} = 1 - a$. Formally, **M** would depend on the frequencies of marker alleles on mutant chromosomes, requiring a larger state space than ancestral/nonancestral alleles. In our case, however, because the recombination rate is much greater than the mutation rate, we assume that the distribution of allele sizes on mutant chromosomes matches that seen in the control population. For simplicity, we also assume a strict stepwise mutation model (Goldstein and Pollock 1997). Under these assumptions, **M** has the elements $M_{11} = 0$, $M_{12} = b/2$, $M_{21} = 1$, and $M_{22} = 1 - b/2$, where b is the frequency of all one-mutant neighbors of the ancestral allele in the control population.

With the parameters of **K** specified, the coalescent time is estimated by multiplying the state vector by **K** iteratively until the observed proportion of ancestral haplotypes is reached. Iteration begins at a frequency vector of (1,0), corresponding to a starting point of only ancestral haplotypes. The analysis requires identification of the ancestral allele, which cannot be determined with certainty when the frequency of the nonancestral type is moderately high, as is the case for both the type II and type III mutations. For the type II mutation, the data are consistent with the ancestral haplotype having carried either the 151- or 153-bp allele, which is consistent with the high frequency of these alleles in the control population. If the 151-bp allele is assumed to have been ancestral, the proportion of ancestral haplotypes among mutant chromosomes (p) is .48, and the frequency of the ancestral allele in the control population (a) is .38, leading to an estimated coalescent time of 120 generations. If the 153-bp allele is assumed to have been

ancestral, $p = .27$ and $a = .22$, and the estimated coalescent time is 189 generations.

The analysis of the type III mutation is complicated by the bimodal frequency distribution among the mutant chromosomes of an allele (145 bp) that is in low frequency in the control populations. In theory, this discrepancy could be caused by any of four factors: (1) multiple origins for the type III mutation; (2) frequencies of the allele in the control population being nonrepresentative of those in the population in which the mutant chromosomes have been recombining; (3) the mutation having occurred originally on the 145-bp allele, despite its rarity in the population; or (4) the occurrence, very early after the appearance of the mutation, of a recombination event between the chromosomes carrying 145- and 151-bp alleles or, equivalently, of a multistep mutation between these alleles. The first explanation is ruled out by the observation of Peretz et al. (1997) that the type III mutation is associated with a single, rare haplotype defined by four closely linked markers. The second explanation also appears unlikely because of the similarity of marker allele frequencies in the Ashkenazi control chromosomes and the CEPH chromosomes (table 1). Although distinguishing the latter two explanations would require more detailed information about the genealogy than is available, we favor the fourth explanation, because it is consistent with the substantially larger differential between the 151- and 153-bp alleles among mutant chromosomes than is observed in the control population. If we accept this explanation, we need not be concerned with whether the mutation actually occurred on a chromosome carrying a 145- or 151-bp allele, but rather we can treat both alleles collectively as ancestral, yielding $p = .78$ and $a = .41$ and leading to an estimated coalescent time of 31 generations. Of course, the coalescent time for the true ancestral haplotype must be longer than this estimate. If we treat the mutation as having occurred singly on either the 145- or 151-bp allele, the estimated age of the mutation would be >100 generations.

Although the expected coalescent times are independent of the shape of the genealogy, confidence intervals are strongly dependent on the precise shape. One approach to the estimation of confidence intervals is to carry out coalescent simulations assuming a range of population growth rates, resulting in gene genealogies ranging from the highly correlated trees typical of constant population size to the star-shaped genealogies typical of very rapid growth. When the data provide information on the number of recombination and mutation events responsible for the observed nonancestral haplotypes, it is possible to select, at least roughly, among these types of genealogies to construct confidence intervals (Reich and Goldstein, in press). Our data set, however, provides relatively little information about the shape of the genealogy, because of the combination of a single marker locus and high frequencies of nonancestral haplotypes. We therefore use a simple heuristic argument to provide a rough guide to the degree of confidence in these estimates, under different assumptions concerning the gene genealogy. For convenience, we shall use the term "confidence interval," but it is important to appreciate that these calculations are meant to be illustrative of how genealogic shape influences confidence and cannot be construed as formal confidence intervals.

First, we assume a star genealogy in which all lineages are uncorrelated—that is, lineages trace their ancestries independently back to the root of the genealogy. In this case, where n is the number of sampled chromosomes and p is the proportion of ancestral chromosomes, there are n independent observations of the time that has elapsed since the appearance of the common ancestor of the sample. In the case of the type II mutation, $n = 99$, $p = .48$, and the confidence interval for the number of ancestral chromosomes observed can be estimated as $\pm 2\times$ the SD of a binomial distribution with parameters .48 and 99, leading to a confidence interval for p of .38–.58. When we apply the recursion in equation (1) to this range, the confidence interval becomes 75–254 generations. A similar calculation for the case of the 153-bp allele having been ancestral leads to a lower bound on the age of the mutation of ~120 generations. An upper bound does not exist in this case, because the lower bound on p is below the equilibrium frequency set by the observed frequency of the ancestral allele in the control population.

The conceptual basis of our approach for assessing confidence intervals in a correlated tree depends on estimating an effective number of lineages that would result in an uncorrelated tree with properties similar to those of a correlated tree with a greater number of sampled chromosomes; that is, we assess the extent to which correlations between lineages in the tree reduce the number of independent observations of the time to the common ancestor. To estimate the effective number of lineages, we determine how many independent lineages would be required to produce a match between the mutant and control chromosome marker allele frequencies at least as close as was actually observed. For example, in the case of the type II mutation, <14 independent lineages would have a <5% chance of leading to a match that is as close as was actually observed.

To better understand the argument, imagine a genealogy in which every chromosome is included within a set of 10 exact copies, but each of these sets traces its ancestry independently to the time of origin. In this case, the number of independent lineages would be 10; however, we would not expect to have an allele-frequency distribution that matches the control distribution as

closely as does the observed distribution. That point is reached only when there are at least 14 independent lineages. Therefore, if we take 14 as the lower bound on the "effective number" of lineages, the youngest possible ages for the type II mutation become 34 and 72 generations for the mutation having occurred on chromosomes carrying the 151- and 153-bp alleles, respectively. In the case of 14 lineages, there is no upper bound on the coalescent time, because the lower bound on p is below the equilibrium value. Applying a similar logic to the type III mutation, we obtain 23 effective lineages, and if we treat both the 145- and 151-bp alleles as ancestral, the estimated confidence interval is 5–70 generations.

This analysis considers uncertainty resulting from the evolutionary process but ignores uncertainty in our estimation of the mutation and recombination rates. Finally, we note that the approach taken here estimates the coalescent time of affected chromosomes, as opposed to the time at which the mutation first appeared, which must predate the coalescent time. Accurate estimation of the date of the mutation would require detailed information about the population's demography at the time of the mutation (see Slatkin and Rannala 1997).

Discussion

The Ashkenazi Jewish population carries a range of single-locus genetic diseases in high frequency (Goodman 1978), and some authors have attributed this to the demographic upheavals thought to have occurred during and shortly after the 16th and 17th centuries (Risch et al. 1995). The factor XI type II mutation, however, has a high frequency in both Iraqi and Ashkenazi Jews. Coupling this observation with the demonstration here that the mutation is old and probably predates the separation of Ashkenazi and Iraqi Jews lends credence to the hypothesis that the frequency of the factor XI type II mutation is largely independent of the recent demographic upheavals particular to the Ashkenazi Jewish population. The frequency of the type II mutation, therefore, seems to require other explanations. Genetic drift in a population ancestral to the major Jewish groups remains a possibility, as does positive selection on heterozygotes. These possibilities may ultimately be distinguishable by comparison of the shape of the genealogy of affected chromosomes with those of the genealogies of apparently neutral genomic regions. Although the severity of the injury-related bleeding phenotype associated with factor XI deficiency depends on nutritional status and the strength of selection against affected individuals is therefore unclear, this analysis suggests the possibility that the frequencies of other genetic disorders common in Ashkenazi Jews may also have little or nothing to do with recent genetic drift in that community. Similar genealogic analyses of other mutations should help to elucidate the forces responsible for their presence in the Ashkenazi population.

Since the best estimate for the coalescent time of the type II mutation is 120 or 185 generations, depending on which allele was ancestral, this analysis also provides evidence that the ancestry of at least some members of both the Iraqi and Ashkenazi Jewish populations can, in fact, be traced to a single ancestral population, presumably residing in the kingdoms of Israel and Judah before the various dispersions. Our findings on the distribution and age of the type II mutation, therefore, mirror recent work on the distribution of the Y chromosomal haplotype characteristic of the Jewish priesthood (Thomas et al. 1998). This haplotype, termed the "Cohen modal haplotype" (CMH), has a very high frequency among self-designated members of the Jewish priesthood and is also found at moderate frequency in lay members of both Sephardic and Ashkenazi Jewry. The CMH, however, has rarely been observed in non-Jewish populations. Finally, whereas the estimation of the age of the factor XI type III mutation is complicated by the unlikely frequency distribution on the mutant chromosomes, this distribution nevertheless appears consistent with a relatively recent origin for that mutation.

References

Bolton-Maggs PHB, Wensley LJ, Tuddenbham EDG (1992) Genetic analysis of 27 kindreds with factor XI deficiency from north west England. Paper presented at 24th Congress of the International Society of Hematology, London, August 23–27, abstract 511a

Goldstein DB, Pollock DD (1997) Launching microsatellites: a review of mutation processes and methods of phylogenetic inference. J Hered 88:335–342

Goodman RM (1978) Genetic disorders among the Jewish people. Johns Hopkins University Press, Baltimore

Peretz H, Mulai A, Usher S, Zivelin A, Segal A, Weisman Z, Mittelman M, et al (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

Reich D, Goldstein DB. Estimating the age of mutations using the variation at linked markers. In: Goldstein DB, Schlötterer C (eds) Microsatellites: evolution and applications. Oxford University Press, Oxford (in press)

Risch N, de Leon D, Ozelius L, Kramer P, Almasy L, Singer B, Fahn S, et al (1995) Genetic analysis of idiopathic torsion dystonia in Ashkenazi Jews and their recent descent from a small founder population. Nat Genet 9:152–159

Saito H, Ratnoff OD, Bouma BN, Seligsohn U (1985) Failure to detect variant (CRM+) plasma thromboplastin antecedent (factor XI) molecules in hereditary plasma thromboplastin antecedent deficiency: a study of 125 patients of several ethnic backgrounds. J Lab Clin Med 106:718–722

Seligsohn U (1978) High gene frequency of factor XI (PTA) deficiency in Ashkenazi Jews. Blood 51:1223

Shpilberg O, Peretz H, Zivelin A, Yatuv R, Chetrit A, Kulka T, Stern C, et al (1995) One of the two common mutations causing factor X1 deficiency in Ashkenazi Jews (type II) is also prevalent in Iraqi Jews, who represent the ancient gene pool of Jews. Blood 85:429–432

Slatkin M, Rannala B (1997) Estimating the age of alleles by use of intraallelic variability. Am J Hum Genet 60:447–458

Stephens JC, Reich DE, Goldstein DB, Doo Shin H, Smith MW, Carrington M, Winkler C, et al (1998) Dating the origin of the *CCR5-32* AIDS resistance allele by the coalescence of haplotypes. Am J Hum Genet 62:1507–1515

Thomas M, Skoreki K, Ben-Ami H, Parfitt T, Bradman N, Goldstein DB (1998) A genetic date for the origin of Old Testament priests. Nature 394:138–140

Weber JL, Wong C (1993) Mutation of human short tandem repeats. Hum Mol Genet 2:1123–1128