Am. J. Hum. Genet. 62:1507–1515, 1998

Dating the Origin of the *CCR5*-Δ*32* AIDS-Resistance Allele by the Coalescence of Haplotypes

J. Claiborne Stephens,[1] David E. Reich,[17] David B. Goldstein,[17] Hyoung Doo Shin,[1] Michael W. Smith,[2] Mary Carrington,[2] Cheryl Winkler,[2] Gavin A. Huttley,[1] Rando Allikmets,[2] Lynn Schriml,[1] Bernard Gerrard,[2] Michael Malasky,[2] Maria D. Ramos,[3] Susanne Morlot,[4] Maria Tzetis,[5] Carole Oddoux,[7] Francesco S. di Giovine,[8] Georgios Nasioulas,[6] David Chandler,[9] Michael Aseev,[10] Matthew Hanson,[1] Luba Kalaydjieva,[9] Damjan Glavac,[11] Paolo Gasparini,[12] E. Kanavakis,[5] Mireille Claustres,[13] Marios Kambouris,[14] Harry Ostrer,[7] Gordon Duff,[8] Vladislav Baranov,[10] Hiljar Sibul,[15] Andres Metspalu,[15] David Goldman,[16] Nick Martin,[18] David Duffy,[18] Jorg Schmidtke,[4] Xavier Estivill,[3] Stephen J. O'Brien,[1] and Michael Dean[1]

[1]Laboratory of Genomic Diversity, and [2]Intramural Research Support Program, Science Applications International Corporation–Frederick, National Cancer Institute, Frederick, MD; [3]Molecular Genetics Department, Hospital Duran i Reynals (IRO), Barcelona; [4]Institut fur Humangenetik, Medizinische Hochschule, Hannover; [5]First Department of Pediatrics, Athens University, St. Sophia's Children's Hospital, and [6]Department of Hygiene and Epidemiology, University of Athens School of Medicine, National Retrovirus Reference Center, Athens; [7]Human Genetics Program, Department of Pediatrics, New York University Medical Center, New York; [8]Department of Molecular and Genetic Medicine, University of Sheffield, Sheffield; [9]Centre for Human Genetics, Edith Cowan University, Perth; [10]Institute of Obstetrics and Gynecology, Russian Academy of Medical Sciences, St. Petersburg; [11]Laboratory of Molecular Pathology, University of Ljubljana, Ljubljana; [12]National Medical Genetics Service, IRCCS-CSS Hospital, San Giovanni Rotondo, Italy; [13]Laboratoire de Biochimie Genetique, CNRS UPR 9008, Montpellier; [14]King Faisal Specialist Hospital and Research Center, Riyadh; [15]Estonian Biocentre and Children's Hospital, University of Tartu, Tartu, Estonia; [16]Laboratory of Neurogenetics, National Institute on Alcohol Abuse and Alcoholism, Rockville; [17]Department of Zoology, University of Oxford, Oxford; and [18]Queensland Institute of Medical Research, Royal Brisband Hospital, Herston, Australia

Summary

The *CCR5*-Δ*32* deletion obliterates the *CCR5* chemokine and the human immunodeficiency virus (HIV)–1 coreceptor on lymphoid cells, leading to strong resistance against HIV-1 infection and AIDS. A genotype survey of 4,166 individuals revealed a cline of *CCR5*-Δ*32* allele frequencies of 0%–14% across Eurasia, whereas the variant is absent among native African, American Indian, and East Asian ethnic groups. Haplotype analysis of 192 Caucasian chromosomes revealed strong linkage disequilibrium between *CCR5* and two microsatellite loci. By use of coalescence theory to interpret modern haplotype genealogy, we estimate the origin of the *CCR5*-Δ*32*–containing ancestral haplotype to be ~700 years ago, with an estimated range of 275–1,875 years. The geographic cline of *CCR5*-Δ*32* frequencies and its recent emergence are consistent with a historic strong selective event (e.g., an epidemic of a pathogen that, like HIV-1, utilizes *CCR5*), driving its frequency upward in ancestral Caucasian populations.

Received December 4, 1997; accepted for publication March 26, 1998; electronically published May 8, 1998.

Address for correspondence and reprints: Dr. Stephen J. O'Brien, Laboratory of Genomic Diversity, National Cancer Institute, Frederick, MD 21702-1201. E-mail: obrien@ncifcrf.gov

0002-9297/98/6206-0030$02.00

Introduction

The *CCR5* gene product encodes a 7-transmembrane G-protein–coupled chemokine receptor that, with CD4, serves as an entry port for primary human immunodeficiency virus (HIV)–1 strains that infect macrophages and monocytes (Alkhatib et al. 1996; Choe et al. 1996; Deng et al. 1996; Doranz et al. 1996; Dragic et al. 1996). In mid-1996, several groups described a 32-bp deletion mutation that interrupts the coding region of the *CCR5* chemokine-receptor locus on human chromosome 3p21 (Dean et al. 1996; Liu et al. 1996; Samson et al. 1996*b*). The *CCR5*-Δ*32* mutation, which leads to truncation and loss of the receptor on lymphoid cells, was remarkable because homozygous individuals had nearly complete resistance to HIV-1 infection despite repeated exposure, and HIV-1 infected heterozygotes for the mutation delay the onset of acquired immunodeficiency syndrome (AIDS) 2–3 years longer than do *CCR5*-+/+ individuals (Dean et al. 1996; Huang et al. 1996; Biti et al. 1997; Michael et al. 1997; O'Brien et al. 1997; Theodorou et al. 1997; Zimmerman et al. 1997). *CCR5*-Δ*32*/Δ*32* homozygotes lack *CCR5*-mediated chemokine responsiveness but do not show immunological pathology, probably because of the genomic redundancy of chemokine-receptor functions (Premack and Schall 1996).

The function-altering nature of the *CCR5*-Δ*32* deletion, a high allele frequency among several Caucasian populations (Dean et al. 1996; Huang et al. 1996; Liu

et al. 1996; Samson et al. 1996*b;* Martinson et al. 1997; Michael et al. 1997), and its rarity or absence in non-Caucasian populations led to speculation that the mutation occurred only once in the ancestry of the Caucasian ethnic group, subsequent to the continental isolation of Caucasians from African ancestors (Dean et al. 1996; O'Brien and Dean 1997). Molecular anthropologists have estimated the date of that separation to be on the order of 200,000 years ago, with a range of 143,000–298,000 years (Cann et al. 1987; Vigilant et al. 1991; Stoneking et al. 1992; Ruvolo et al. 1993; Goldstein et al. 1995; Horai et al. 1995; von Haeseler et al. 1996). Furthermore, in attempts to explore the veracity of the mitochondrial "Eve Hypothesis," considerable evidence has been assembled that argues against the occurrence of a significant population bottleneck or demographic contraction since that early divergence of ethnic group ancestors (Takahata et al. 1992; Ayala 1995; Ayala and Escalante 1996). In fact, several estimates of prehistoric (15,000–200,000 years ago) population sizes of humans have converged as 10,000–100,000 individuals (Takahata et al. 1992; Ayala 1995; Ayala and Escalante 1996). For such large populations, it is well established that new mutations would have a very high likelihood (>95%) of being lost within a few dozen generations (Fisher 1930; Kimura and Ohta 1971). It is not impossible, albeit highly unlikely, that a single *CCR5*-Δ*32* variant did increase to modern frequencies across Europe/Asia, by random genetic drift, as a strictly neutral mutation.

In this report, we present a new survey of *CCR5*-Δ*32* allele frequency in 38 ethnic populations including 4,166 individuals (table 1). A north-to-south cline of allele frequency is affirmed (Martinson et al. 1997; Libert et al. 1998) as well as the absence of *CCR5*-Δ*32* among East Asian, Middle Eastern, and American Indian populations. The time of origin of the *CCR5*-Δ*32* mutation was estimated on the basis of the persistence of a common and likely ancestral three-locus haplotype (including *CCR5*-Δ*32* and specific alleles of two adjacent microsatellite loci) retained in linkage disequilibrium across 0.9 cM on chromosome 3, among modern *CCR5*-Δ*32*–bearing chromosomes. We suggest that this most common ancestral *CCR5*-Δ*32*–bearing haplotype ($P = .85$) arose by a unique deletion mutation of the *CCR5*-+ allele on the most common *CCR5*-+ haplotype ($P = .36$) and that this haplotype was elevated by natural selective pressures (likely on the *CCR5*-Δ*32* allele) to present frequencies of 5%–15%. Following the selective increase, derivative modern Caucasian haplotypes appeared, allowing a coalescence-based estimation of the time required to produce the present haplotype distribution. The age of that *CCR5*-Δ*32*–bearing haplotype and possibly the *CCR5*-Δ*32* variant was computed, by use of a Markov expansion, as ~700 years old (range 275–1,875 years).

Table 1

Frequency of the *CCR5*-Δ*32* Allele in Defined Populations, Ranked in Descending Order of Δ32 Frequency

Ethnic Group	No. of Individuals	Allele Frequency	SD
Swedish	131	.137	.021
Russian	50	.136	.034
Estonian	158	.133	.019
Polish	30	.133	.044
Slovakian	30	.133	.044
Tatar	50	.120	.032
Australian	395	.118	.011
British	422	.117	.011
Irish	31	.113	.040
German	208	.108	.015
Czech	161	.102	.017
Spanish	56	.098	.028
Ashkenazi	503	.097	.009
Finn	195	.091	.015
French	230	.089	.013
Austrian	36	.089	.033
Danish	24	.083	.040
Albanian	73	.082	.023
Slovenian	110	.077	.018
Turkish	40	.063	.027
Italian	172	.055	.012
Azerbaijani	40	.050	.024
Bulgarian	29	.045	.027
Greek	160	.044	.011
Uzbek	29	.034	.024
Bulgarian Gypsy	47	.032	.018
Kazakh	50	.030	.017
Mexican	42	.024	.017
Uigur	45	.022	.016
Tuvinian	50	.020	.014
Georgian	50	.00	.00
Lebanese	51	.00	.00
Saudi	100	.00	.00
Cheyenne	100	.00	.00
Pima Indian	78	.00	.00
Pueblo Indian	100	.00	.00
Korean	50	.00	.00
Chinese	40	.00	.00

NOTE.—Population allele frequency SDs were estimated by assuming that allele frequencies are binomially distributed—that is, $SD = \sqrt{pq/2n}$, where n is the sample size for each population. All population genotype frequencies conformed to Hardy-Weinberg equilibrium.

Methods

Radiation-Hybrid Mapping

Primers for the *CCR1, CCR4, GAAT12D11, AFMB362wb9, STRL33, D3S3582,* and *D3S3647* markers were used to type the GeneBridge 4 panel of radiation hybrids (Research Genetics) to determine centiray position, and data were submitted to the radiation-hybrid mapping service at the Whitehead Institute.

Haplotype Analysis

Individuals homozygous for *CCR5*-Δ*32* and CEPH families carrying the *CCR5*-Δ*32* allele were used to determine chromosomal haplotype phase for variants at *CCR5* and seven microsatellite loci. Pairwise tests between loci revealed strong linkage disequilibrium between *CCR5*-Δ*32* and two flanking short tandem-repeat polymorphic (STRP) markers, *GAAT12D11* (197-bp allele) and *AFMB362wb9* (215-bp allele). We abbreviate these loci as *GAAT* and *AFMB,* respectively.

Age Estimation, Based on Current Frequency

The average age of a neutral two-allele polymorphism with frequencies p and $1 - p$ is $-4N_e[p(\log_e p) + (1 - p)\log_e(1 - p)]$ (Kimura and Ohta 1973), which yields 6,500 generations for *CCR5*-Δ*32*, on the basis of the assumption of $p = .10$ and $N_e = 5{,}000$ for Caucasians. Under the assumption of 25 years per human generation, the age of the polymorphism would be estimated to be 162,500 years. This estimate is likely inappropriate, since it is based on two scenarios weighted by the probability of their occurrence: that of the *CCR5*-Δ*32* allele rising from nearly 0 to its current frequency p and that of its dropping from near fixation to p (Kimura and Ohta 1973). Since the *CCR5*-Δ*32* mutation is absent in all East Asian and African populations tested, it seems to have a more recent origin than the wild type, so a better estimate is $-4N_e[p(\log_e p)/(1 - p)]$ (Kimura and Ohta 1973), which yields 5,100 generations, on the basis of the assumption of $N_e = 5{,}000$ for Caucasians. Under the assumption of selective neutrality, genetic drift, and 25 years per human generation, the age of the *CCR5*-Δ*32* mutation would now be estimated to be 127,500 years.

Age Estimation, Based on Interhaplotype Variation

This method considers the chromosomal haplotypes defined by STRP loci in linkage disequilibrium with *CCR5*-Δ*32* as indicators of derivative events for which we can estimate the frequency on the basis of mutation and recombination rates (see Kaplan et al. 1994; Risch et al. 1995; Tishkoff et al. 1996). First, we will identify the most likely ancestral *CCR5*-Δ*32* haplotype and then estimate the proportion of *CCR5*-Δ*32* haplotypes that exhibit no change from the ancestral haplotype. Assuming that mutation and recombination occur at a combined rate r, we can then use the proportion of unchanged haplotypes to estimate the age of origin.

The probability P that a given haplotype does not change from its ancestor G generations ago is simply

$$P = (1 - r)^G \approx e^{-rG} . \quad (1)$$

To estimate P, we note that for a dramatically expanded population—one for which all lineages are essentially independent—an unbiased estimate of P is the proportion of observed haplotypes that are ancestral (Risch et al. 1995). Although at first surprising, this also holds true for a constant-sized population in which many lineages are highly correlated, in the sense that pairs of alleles share extensive periods of coancestry during the time tracing back to the most recent common ancestor of the sample. The age estimate is independent of topology because, as long as mutations at the marker loci have no selective effect, the correlations in the tree amount to a process of pseudoreplication of lineages (Reich et al., in press). This process will affect the variance of our estimate of P; however, because the lineages that are replicated are not subject to selection for allelic state, the proportion of ancestral haplotypes will not be systematically affected (Reich et al., in press).

Using this approach, we can easily estimate G in terms of P. In particular, by transforming equation (1) to

$$G = -\ln(P)/r , \quad (2)$$

we obtain an unbiased estimate of the age of the most recent common ancestor of the sampled haplotypes. Although this estimate is not affected by tree topology, the variance of the age estimate depends strongly on the shape of the tree (Reich et al., in press). For a tree with highly correlated lineages (typical of a constant-sized population [Slatkin and Hudson 1991]), the variance will tend to be relatively large because there are few independent samplings of the age of the tree. In contrast, for the starlike topology typical of an expanding population, the variance will be smaller because the sample represents more independent observations. Note that the amount of correlation in the tree can be assessed directly from the distribution of nonancestral haplotypes, and such information can be incorporated into computer simulations used to estimate the variance (Reich et al., in press). Knowledge of historical population sizes can also be used to constrain the date estimate (see Results).

Estimation of r

We need to estimate r to use the preceding theory. Although we do not have mutation-rate estimates specifically for *GAAT* and *AFMB,* Weber and Wong (1993) have estimated rates for a large number of microsatellite loci. From these, we will assume a rate of $\mu = .001$ as an upper limit for mutation at either locus. To justify this, we note that the number of alleles at *GAAT* ($n = 3$) and at *AFMB* ($n = 4$) are relatively small compared with the range (6–17) seen in our other sampled microsatellite loci. Next, we require the recombination rate (c) among *CCR5, GAAT,* and *AFMB*. Although we do not have direct estimates of recombination, these loci have been ordered physically using a radiation-hybrid

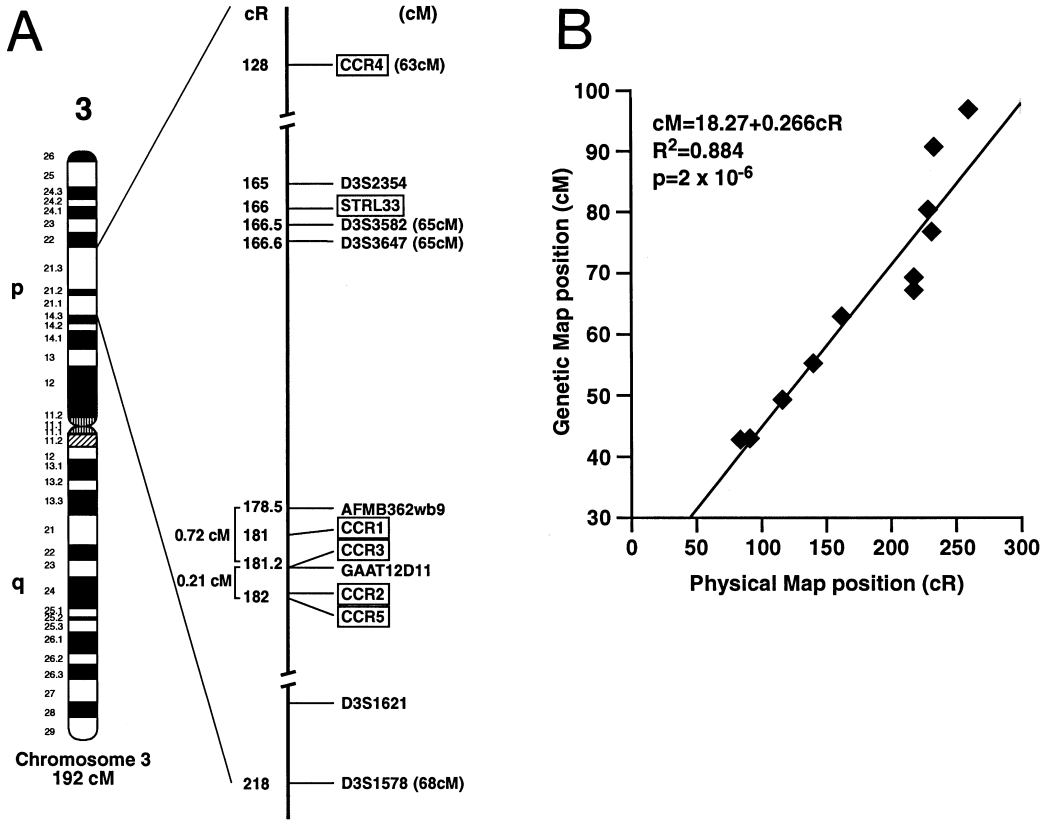

Figure 1 *A,* Map of the chromosome 3p21 region containing the *CCR* gene complex. The position of the chemokine-receptor genes on chromosome 3p is shown in relation to neighboring microsatellite markers. The position of the genes and markers is shown on the physical map produced by radiation-hybrid analysis, and distances are given in centirays. In parentheses are centimorgan positions based on recombination for CEPH families (Dib et al. 1996). The *CCR1, CCR2, CCR3,* and *CCR5* genes have been shown to reside within 300 kb of each other (Raport et al. 1996; Samson et al. 1996*a*). Analyses of genetic and physical distances in this region indicate that 1 cM is equivalent to ~3.76 cR. Radiation-hybrid–map positions are from the Whitehead Institute or were determined in this study. *B,* Regression of recombination distance (cM) versus physical distance (cR) of 13 STRP loci on chromosome 3 for which both centimorgan and centiray data were available (Dib et al. 1996; G. A. Huttley, unpublished data). STRP loci, examined in linear centiray order, were *D3S1567, 1583, 1609, 1561, 1611, 3564, 1588, 1582, 1578, 1312, 1313, 1285,* and *1566.*

map (fig. 1*A*), and distances have been estimated. From a regression of STRP loci on chromosome 3 (fig. 1*B*), we obtain the conversion 1 cM = 3.76 cR. Present frequencies of the different wild-type haplotypes (table 2) were used to infer the fraction of recombination events that result in the *CCR5-Δ32* mutation on non–*197-215* haplotypes.

From the map, *CCR5*–(0.8 cR)–*GAAT*–(2.7 cR)–*AFMB,* we estimate 0.21% recombination between *CCR5* and *GAAT* and 0.72% between *GAAT* and *AFMB*. In the first case, 93 (64%) of 146 *CCR5*-+–containing haplotypes are not +-*197-215*, so that ~2/3 of the recombination events between *CCR5-Δ32*–bearing and *CCR5*-+ haplotypes would result in transfer of *CCR5-Δ32* to a different haplotype. In the second case, 70 (48%) of 146 *CCR5*-+ haplotypes do not have the *AFMB-215* allele, and hence almost half result in observed recombination. Combining these, c = .64 (0.21%) + .48(0.72%) = .005 is our estimate of the rate of recombination events involving the *CCR5-Δ32-197-215* haplotype that actually lead to transfer of the *CCR5-Δ32* mutation to a different haplotype. We then combine this estimate with μ = .001 above, for mutation, to get r = .006 as our estimate of the total rate of change from either mutation or recombination. This calculation does not consider regeneration of the ancestral *CCR5-Δ32*–bearing haplotype by recombination, because this value is neglible (see Results).

Estimation of Selective Coefficients

To calculate the magnitude of selection needed to increase the frequency of *CCR5-Δ32* from essentially 0% to 10%, in G generations, we set up an iteration using the standard equation for gene-frequency change under selection $p' = p(pw_{11} + qw_{12})/\overline{w}$, in which the trio w_{11},

Table 2

***CCR5* Haplotypes Observed in Modern Caucasians**

Haplotype	*N* (%)
*CCR5*Δ*-32*	
CCR5-GAAT-AFMB:	
Δ*32-197-215*[a]	39 (84.8)
Δ*32-197-217*[b]	3 (6.5)
Δ*32-193-215*[b]	2 (4.3)
Δ*32-197-219*[c]	1 (2.2)
Δ*32-197-213*[d]	1 (2.2)
Total	46 (100)
CCR5-+	
CCR5-GAAT-AFMB:	
+-197-215	53 (36.3)
+-197-217	45 (30.8)
+-193-215	20 (13.7)
+-197-219	2 (1.4)
+-193-217	21 (14.4)
+-191-217	2 (1.4)
+-191-215	3 (2.1)
Total	146 (100)

[a] Ancestral haplotype.
[b] Recombinational origin.
[c] Either mutational or recombinational origin.
[d] Mutational origin.

w_{12}, and $\overline{w}$ is adjusted depending on whether *CCR5*-Δ*32* is dominant, codominant, or recessive to wild type (Hartl and Clark 1989). For example, if *CCR5*-Δ*32* is dominant, $w_{11} = w_{12} = 1$, $w_{22} = 1 - s$, and $\overline{w} = 1 - sq^2$, so that $p' = spq^2$. Trial values of s are increased until p' becomes 10% after G generations of selection. Initial values of p were .0005 and .0001, corresponding to $p = 1/2N_e$ for $N_e = 1{,}000$ and 5,000, respectively.

Results

Genomic DNA samples obtained from 4,166 individuals, as identified in 38 ethnic groups from Europe, Asia, the Middle East, and North America, were typed for *CCR5* (Dean et al. 1996). The results (table 1) suggest a north-to-south gene-frequency gradient (or cline), with the highest allele frequencies in northern Europe (14%) to a low of 4.4% in Greece. The *CCR5*-Δ*32* allele was not found among Lebanese, Georgian, Saudi, Korean, Chinese, or American Indian (Cheyenne, Pueblo, and Pima) populations in samples of 40–100 individuals. However, significant frequencies of the allele were apparent among Central Asian groups such as Azerbaijanis, Uigurs, Uzbeks, Kazakhs, Tuvinians, and Tatars. These data confirm the high frequency of *CCR5*-Δ*32* among northern European Caucasians, a gene-frequency cline across Europe and Asia reflecting recent population admixture, and virtual absence of *CCR5*-Δ*32* among native Africans, East Asians, and American Indians (Dean et al. 1996; Huang et al. 1996; Michael et al. 1997; Martinson et al. 1997; Libert et al. 1998).

In order to estimate the time interval that elapsed since the occurrence of the *CCR5*-Δ*32* mutation, we examined the disposition of polymorphic loci adjacent to *CCR5*, in modern Caucasian populations. The *CCR5* locus has been mapped to chromosome 3p21 and was found to be tightly linked to at least four other genetically homologous CC-chemokine–receptor (*CCR*) genes, *CCR1–4* (Combadiere et al. 1996; Dean et al. 1996; Samson et al. 1996*a*). Adjacent to the *CCR* genes are seven STRP loci. We have determined the physical order of the *CCR* and STRP loci using a radiation-hybrid panel (fig. 1*A*). Physical centiray distances were converted to recombination distances (in centimorgans) by use of a regression of centiray versus centimorgan distances computed for 13 STRP loci mapped to chromosome 3 by use of both linkage and radiation hybrids (fig. 1*B*).

In order to examine composite *CCR5* allele–containing haplotypes, we genotyped 19 *CCR5*-Δ*32*/Δ*32* homozygotes and 72 *CCR5*-+/+ homozygotes from AIDS cohorts (Dean et al. 1996) plus 20 *CCR5*-+/Δ*32* heterozygotes and 17 *CCR5*-+/+ homozygotes from the CEPH mapping families, for the seven adjacent STRP loci (fig. 1*A*). Linkage disequilibrium was tested for all independent phase-known locus pairs and was strongly evident for *CCR5* and the two STRP loci nearest to *CCR5* (*GAAT12D11* and *AFMB362wb9*). These loci were mapped by use of radiation hybrids, and their recombination interval was estimated, from figure 1*B,* as 0.21 and 0.93 cM, respectively, from *CCR5*. A common (p = 10%–15%) missense-mutation allele (64I) of the *CCR2* locus 18 kb from *CCR5* is also in complete linkage disequilibrium with *CCR5*-+ (Smith et al. 1997). Other STRP loci in the region (fig. 1*A*) that are at greater linkage distances (>4.1 cM) show lower or no level of linkage disequilibrium with *CCR5*. High linkage disequilibrium of *CCR5*-Δ*32* with two adjacent STRP loci is consistent with the *CCR5* deletion mutation descending from a unique mutation in recent history.

In table 2, we list the composite three-locus haplotype of five *CCR5*-Δ*32*–containing and seven *CCR5*-+–containing haplotypes and their frequency among 192 phase-known chromosomes typed in our sample. The nonrandom association of STRP and *CCR5* alleles, their most parsimonious phylogenetic history, and present haplotype frequencies were used to calculate the time required for a new mutation of an ancestral haplotype to produce the modern distribution of haplotypes, on the basis of coalescent theory (Hudson and Kaplan 1986; Hudson 1990). That is, the development of new haplotypes measurable in modern populations (table 2) reflects accumulation of mutational and recombina-

tional evolution of the ancestral haplotype since its origin or selective elevation in ancestral populations.

To use the above theory, we note that 39 of 46 *CCR5-Δ32*–bearing haplotypes were identical: *Δ32-197-215*; the four additional *CCR5-Δ32*–bearing haplotypes included different STRP alleles (table 2). Among the *CCR5-Δ32*–bearing haplotypes, *Δ32-197-215* is by far the most frequent haplotype (84.8%) and is a single mutation step from the most common *CCR5-+* haplotype, *+-197-215* (table 2). Thus, we may assume that *CCR5-+-197-215* is the ancestral haplotype on which the *CCR5-Δ32* mutation arose (Watterson and Guess 1977) and that the other *CCR5-Δ32*–bearing haplotypes were derived from it by four to seven mutational or recombinational events. Substituting the present frequency (.848) of the ancestral *CCR5-Δ32*–bearing haplotype for *P,* the probability that a given haplotype is unchanged from its ancestor, in equation (2) with our estimate of *r* (.006, the rate of combined mutational/recombinational change of that haplotype; see Methods), we obtained an estimate of 27.5 generations, or 688 years, for the origin and expansion of the *CCR5-Δ32* ancestral haplotype, on the basis of a 25-year human-generation time.

Two potential sources of error in our estimates of *r* (mutation/recombination frequency) and *p* (the frequency of ancestral haplotype) deserve comment. First, our estimates are sensitive to *r,* which itself is dominated by recombination, since the estimated recombination rates are several-fold greater than the estimated STRP mutation rate (see Methods). Our estimated *r* value is based on a regression of centirays versus centimorgans (fig. 1*B*), using 13 STRP loci mapped on chromosome 3, with both linkage and radiation hybrids. The regression shows a high precision or correlation ($r^2 = .884$; $p = 2 \times 10^{-6}$; fig. 1*B*), although there is a modest departure in the centimorgan:centiray concordance in the actual region (175–185 cR) where the haplotype resides (see fig. 1*B*), suggesting a 10%–20% reduction in recombination for that region. If we consider lower *r* values (e.g., $r = .004$ or .002) the *G* estimates become 41.3 generations (1,032 years) and 82.5 generations (2,064 years), respectively, which still are within the range of recorded human history. (An extremely conservative computation, calibrating the D3S3647–D3S1578 distances at 3 cM and 51.4 cR, reflecting an apparent but still uncertain reduction in recombination over the *CCR* cluster [see fig. 1*A*], yields an estimate of $r = .002$, or 2,064 years for the haplotype age).

Variance of the estimate of coalescence time *G* due to variability of our ancestral haplotype-frequency estimate ($p = .848$) was addressed by determining the frequency of derived or nonancestral two-locus haplotypes (i.e., not *CCR5-Δ32-197-X* or *CCR5-Δ32-X-215*, where *X* is undetermined; see table 2) in a group of 1,400 chromosomes. The sampling revealed frequencies of 9.2% for *CCR5-Δ32-193-X* and 7.6% for *CCR5-Δ32-X-217* plus *CCR5-Δ32-X-219,* which sums to 16.8% nonancestral haplotypes, remarkably close to the 15.2% nonancestral haplotypes determined for nonancestral three-locus haplotypes in our sample (table 2). Substituting 9.2% and 16.8% as lower and upper limits of derived haplotype frequencies, we computed (equation [2]) an alternative estimate of *G* equal to 16–31 generations (402–766 years) as an indication of the influence of sampled haplotype frequency on *G*.

The general coalescence prediction of $G = 28$ generations was examined empirically by incorporating a complete Markov transition matrix into a computer simulation based on a coalescent algorithm (Hudson 1990). This approach considers regeneration of the ancestral haplotype and assesses confidence intervals for a range of possible growth models (and hence range of degrees of correlation in the genealogy coalescence) (Reich et al., in press). We performed 1,000 simulations for each combination of demographic parameters, for population sizes $\leqslant$100,000 and for exponential growth rates from zero to rapid growth, and found that only a narrow range of demographic parameters were consistent with the observed number and distribution of nonancestral haplotypes. By using the variance of the time depth of the simulated trees for the combinations of demographic parameters that were allowed and by making the further assumption that European population sizes during the past several thousand years have been moderately large ($N > 5{,}000$), we were able to restrict the range of allowable dates (95% confidence interval) to 11–75 generations (or 275–1,875 years) ago.

Discussion

The data reported here and elsewhere (Ansari-Lari et al. 1997; Carrington et al. 1997; Martinson et al. 1997; O'Brien and Dean 1997; Libert et al. 1998) provide indirect but persuasive evidence for the recent unique occurrence of a deletion mutation in the *CCR5* locus that mediates host response to HIV. The *CCR5-Δ32* allele, which leads to abolishment of the *CCR5* function, occurs exclusively among Caucasians and describes a north-to-south geographic cline with a high frequency of 14% among Swedes to 5% among Mediterranean peoples to 0% among Saudi and East Asian populations. The *CCR5-Δ32* allele is retained in a 0.9-cM haplotype on chromosome 3 that has persisted in linkage disequilibrium in human populations for ~700 years.

The recency of occurrence plus the key role played by *CCR5* as a requisite coreceptor for both HIV-1 infection and progression to AIDS (Dean et al. 1996; Huang et al. 1996; Liu et al. 1996; Samson et al. 1996*b;* Zimmerman et al. 1997) leads to the suggestion that a strong

selective pressure, such as a widespread fatal epidemic, should be invoked to explain the allele-frequency distribution observed in modern Eurasia. The selective hypothesis targeting *CCR5* draws further support from (1) the absence of clinical or immunological pathology among *CCR5-Δ32/Δ32* homozygotes, in spite of their complete loss of chemokine-receptor function (Dean et al. 1996; Liu et al. 1996) and (2) a recent demonstration that 14 (81%) of 17 naturally occurring *CCR5* mutations were codon altering or nonsynonymous (Carrington et al. 1997). This level of nonsynonymous substitution is far greater than the frequency seen in a sequence comparison of 49 human genes with their mouse homologues (15% nonsynonymous [Li 1997]). Elevated numbers of nonsynonymous substitutions are generally interpreted as evidence for selective pressure for amino-acid–sequence divergence, such as is observed in the mammalian major histocompatibility complex (Hughes and Nei 1988).

The coalescence-based estimate, which is supported by simulation analysis (Reich et al., in press), places the origin of the *CCR5-Δ32-197-215* haplotype in very recent historic times, in marked contrast with the date computed under a strictly neutral genetic-drift model (127,500 years; see Methods). The disparity in the two estimates also would be explained by a strong selective pressure favoring the *CCR5-Δ32*–bearing haplotype and perhaps mediated by the *CCR5-Δ32*–specified phenotype, during human history.

The high allele frequency of a number of hereditary recessive diseases in specific outbred populations has been explained by a heterozygote advantage of the mutant allele that could compensate for the deleterious effect of homozygotes. The best-known example is the connection between sickle-cell anemia, thalassemia, and Duffy mutations balanced by malaria resistance (Chaudhuri et al. 1995; Gelpi and King 1976; Vogel and Motulsky 1997). Similar hypotheses for the frequency of Tay-Sachs disease and cystic fibrosis have been proposed (O'Brien 1991; Gabriel et al. 1994; Morral et al. 1994; Macek et al. 1997). Although it is possible for genetic drift to cause an individual allele to reach an elevated frequency, the probability of this occurring very rapidly is minuscule in large outbred groups (Fisher 1930; Kimura and Ohta 1971). For instance, the probability of a new mutation reaching 10% within 28 generations by drift alone is 6.2×10^{-8}, on the assumption that it starts at $1/2N_e$, with $N_e = 1{,}000$. Recurrent mutation and/or selection are potential alternative explanations for the high frequency of *CCR5-Δ32* in Europe. However, recurrent mutation is unlikely, since the *CCR5-Δ32* allele was not found in African or East Asian groups and occurs largely in a homogeneous haplotypic background (table 2).

Deterministic models are appropriate for exploring the apparent rapidity of gene-frequency change that selection is postulated to mediate. Positive selection coefficients of 23% (dominance) or 37% (additivity), favoring the *CCR5-Δ32*–positive allele, would have been required, to increase the frequency from 1/10,000 to 10% within 28 generations (Hartl and Clark 1989). For smaller selection coefficients, even more generations would be required. Completely recessive alleles would require enormous selection coefficients, even for 5,000 generations. The sum of these considerations provides considerable, albeit indirect, support for the scenario that the *CCR5-Δ32* mutation occurred once, on the order of 700 years ago, in a Caucasian population, and has rapidly increased in its frequency by a strong selective pressure, possibly an ancient plague, the nature of which is currently undetermined.

The estimates derived here track the persistance of the three-locus haplotype at 700 years; however, it is possible that the *CCR5-Δ32* mutation is somewhat older, particularly if multiple pulses of selective episodes on *CCR5-Δ32* were involved. In spite of that uncertainty, the cumulative results point to a selective sweep and to one with enormous selective mortality within historic times, perhaps mediated by a widespread epidemic. The bubonic plague, which claimed the lives of 25%–33% of Europeans during the Black Death from 1346 to 1352 (650 years ago) and which has had multiple outbreaks in Europe before and since, is an obvious candidate (Lenski 1988; McEvedy 1988). The plague bacillus, *Yersinia pestis,* is transmitted by fleas on black rats and carries a 70-kb plasmid (PYV), which encodes an effector protein, Yop1, that enters macrophages, causing diminished immune defenses (Rosqvist et al. 1988; Cornelis and Wolf-Wulz 1997; Mills et al. 1997). If the mechanism of *Yersinia*-induced macrophage apoptosis (cell death) involved macrophage chemokine receptor 5, the *CCR5-Δ32* mutation would be an attractive candidate for a strong selective pressure 600–700 years ago. Other possibilities are *Shigella, Salmonella,* and *Mycobacterium tuberculosis,* which likewise target macrophages. Additional infectious-disease candidates would include syphilis, small pox, and influenza, which have decimated millions of individuals during the previous millennium (McNeil 1976; Garrett 1994). Attempts to examine these deadly pathogens of documented mortality during the dawn of Western civilization, in the context of the *CCR5* genotype, would be illuminating.

Acknowledgments

We thank Teri Kissner, Raleigh Boaze, Janine Timms, Carol Mayne, and Stan Cevario, for technical assistance. Computing resources were provided by the Frederick Biomedical Supercomputing Center. We would also like to thank Drs. Michael Clegg and Bruce Weir, for reviewing an early version of this

manuscript, and Dr. Al Tolun (Bogazici University, Istanbul), for providing the Turkish samples.

Electronic-Database Information

URLs for data in this article are as follows:

Whitehead Institute, http://www-genome.wi.mit.edu

References

Alkhatib G, Combadiere C, Broder CC, Feng Y, Kennedy PE, Murphy PM, Berger EA, et al (1996) CC CKR5: a RANTES, MIP1α, MIP-1β receptor as a fusion cofactor for macrophage-tropic HIV-1. Science 272:1955–1958

Ansari-Lari MA, Liu X-M, Metzker ML, Rut AR, Gibbs RA (1997) The extent of genetic variation in the CCR5 gene. Nat Genet 16:221–222

Ayala FJ (1995) The myth of Eve: molecular biology and human origins. Science 270:1930–1936

Ayala FJ, Escalante AA (1996) The evolution of human populations: a molecular perspective. Mol Phylogenet Evol 5: 188–201

Biti R, French R, Young J, Bennetts B, Stewart G (1997) HIV-1 infection in an individual homozygous for the CCR5 deletion allele. Nat Med 3:252–253

Cann RL, Stoneking M, Wilson AC (1987) Mitochondrial DNA and human populations. Nature 325:31–36

Carrington M, Kissner T, Gerrard B, Ivanov S, O'Brien SJ, Dean M (1997) Novel allele of the chemokine-receptor gene *CCR5*. Am J Hum Genet 61:1261–1267

Chaudhuri A, Polyakova J, Zbrzezna V, Pogo AO (1995) The coding sequence of Duffy blood group gene in humans and simians: restriction fragment length polymorphism, antibody and malarial parasite specificities, and expression in nonerythroid tissues in Duffy-negative individuals. Blood 85:615–621

Choe H, Farzan M, Sun Y, Sullivan N, Rollins B, Ponath PD, Wu L, et al (1996) The β-chemokine receptors CCR3 and CCR5 facilitate infection by primary HIV-1 isolates. Cell 85:1135–1148

Combadiere C, Ahuja SK, Tiffany HL, Murphy PM (1996) Cloning and functional expression of CC CKR5, a human monocyte CC chemokine receptor selective for MIP-1(α), MIP-1(β), and RANTES. J Leukoc Biol 60:147–152

Cornelis GR, Wolf-Watz H (1997) The Yersinia Yop virulon: a bacterial system for subverting eukaryotic cells. Mol Microbiol 23:861–867

Dean M, Carrington M, Winkler C, Huttley GA, Smith MW, Allikmets R, Goedert JJ, et al (1996) Genetic restriction of HIV-1 infection and progression to AIDS by a deletion allele of the CKR5 structural gene. Science 273:1856–1862

Deng H, Liu R, Ellmeier W, Choe S, Unutmaz D, Burkhart M, Di Marzio P, et al (1996) Identification of a major co-receptor for primary isolates of HIV-1. Nature 381:661–666

Dib C, Fauré S, Fizames C, Samson D, Drouot N, Vignal A, Millasseau P, et al (1996) A comprehensive genetic map of the human genome based on 5,264 microsatellites. Nature 380:152–154

Doranz BJ, Rucker J, Yi Y, Smyth RJ, Samson M, Peiper S, Parmentier M, et al (1996) A dual-tropic primary HIV-1 isolate that uses fusin and the β-chemokine receptors CKR-5, CKR-3, and CKR-2b as fusion cofactors. Cell 85: 1149–1158

Dragic T, Litwin V, Allaway GP, Martin SR, Huang Y, Nagashima KA, Cayanan C, et al (1996) HIV-1 entry into CD4+ cells is mediated by the chemokine receptor CC-CKR5. Nature 381:667–673

Fisher RA (1930) The distribution of gene ratios for rare mutations. Proc R Soc Edinburgh 50:205–220

Gabriel SE, Brigman KN, Koller BH, Boucher RC, Stutts MJ (1994) Cystic fibrosis heterozygote resistance to cholera toxin in the cystic fibrosis mouse model. Science 266: 107–109

Garrett L (1994) The coming plague. Penguin, New York

Gelpi AP, King MC (1976) Association of Duffy blood groups with sickle cell trait. Hum Genet 32:65–68

Goldstein D, Ruiz Linares A, Cavalli-Sforza LL, Feldman MW (1995) Genetic absolute dating based on microsatellites and the origin of modern humans. Proc Natl Acad Sci USA 92: 6723–6727

Hartl DL, Clark AG (1989) Principles of population genetics. Sinauer, Sunderland, MA

Horai S, Hayasaka K, Kondo R, Tsugane K, Takahata N (1995) Recent African origin of modern humans revealed by complete sequences of hominoid mitochondrial DNAs. Proc Natl Acad Sci USA 92:532–536

Huang Y, Paxton WA, Wolinsky SM, Neumann AU, Zhang L, He T, Kang S, et al (1996) The role of a mutant CCR5 allele in HIV-1 transmission and disease progression. Nat Med 2:1240–1243

Hudson RR (1990) Gene genealogies and the coalescent process. Oxf Surv Evol Biol 7:1–44

Hudson RR, Kaplan NL (1986) On the divergence of alleles in nested subsamples from finite populations. Genetics 113: 1057–1076

Hughes AL, Nei M (1988) Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. Nature 335:167–170

Kaplan NL, Lewis PO, Weir BS (1994) Age of the ΔF508 cystic fibrosis mutation. Nat Genet 8:216–217

Kimura M, Ohta T (1971) Theoretical aspects of population genetics. Princeton University Press, Princeton, NJ

——— (1973) The age of a neutral mutant persisting in a finite population. Genetics 75:199–212

Lenski RE (1988) Evolution of plague virulence. Nature 334: 473–474

Li WH (1997) Molecular evolution. Sinauer, Sunderland, MA

Libert F, Cochaus P, Beckman G, Samson M, Aksenova M, Cao A, Czeizel A, et al (1998) The ΔCCR5 mutation conferring protection against HIV-1 in Caucasian populations has a single and recent origin in northeastern Europe. Hum Mol Genet 7:399–406

Liu R, Paxton WA, Choe S, Ceradini D, Martin SR, Horuk R, MacDonald ME, et al (1996) Homozygous defect in HIV-1 coreceptor accounts for resistance of some multiply-exposed individuals to HIV-1 infection. Cell 86:367–377

Macek M Jr, Macek M Sr, Krebsova A, Nash E, Hamosh A, Reis A, Varon-Mateeva R, et al (1997) Possible association of the allele status of the CS.7/*Hha*I polymorphism 5′ of the

CFTR gene with postnatal female survival. Hum Genet 99: 565–572

Martinson JJ, Chapman NH, Rees DC, Liu Y-T, Clegg JB (1997) Global distribution of the CCR5 gene 32–base pair deletion. Nat Genet 16:100–102

McEvedy C (1988) The bubonic plague. Sci Am 258:118–123

McNeil WH (1976) Plagues and people. Blackwell, London

Michael NL, Chang G, Louie LG, Mascola JR, Dondero D, Birx DL, Sheppard HW (1997) The role of viral phenotype and CCR-5 gene defects in HIV-1 transmission and disease progression. Nat Med 3:338–340

Mills SD, Boland A, Sory MP, Van De Smissen P, Kerbourch C, Finlay BB, Cornelis GR (1997) *Yersinia enterocolitica* induces apoptosis in macrophages by a process requiring functional type III secretion and translocation mechanisms and involving YopP, presumably acting as an effector protein. Proc Natl Acad Sci USA 94:12638–12643

Morral N, Bertranpetit J, Estivill X, Nunes V, Casals T, Gimenez J, Reis A, et al (1994) The origin of the major cystic fibrosis mutation (ΔF508) in European populations. Nat Genet 7:169–175

O'Brien SJ (1991) Ghetto legacy: can the high incidence of Tay-Sachs disease in Ashkenazi Jews be linked to historic epidemics of tuberculosis in industrial European cities? Curr Biol 1:209–211

O'Brien SJ, Dean M (1997) In search of AIDS-resistance genes. Sci Am 277:44–51

O'Brien T, Winkler C, Dean M, Nelson JAE, Carrington M, Michael NL, White GC II (1997) HIV-1 infection in a man homozygous for CCR5-Δ32. Lancet 349:1219

Premack BA, Schall TJ (1996) Chemokine receptors: gateways to inflammation and infection. Nat Med 2:1174–1178

Raport CJ, Gosling J, Schweickert VL, Gray PW, Charo IF (1996) Molecular cloning and functional characterization of novel human CC chemokine receptor (CCR5) for RANTES, MIP-1β, and MIP-1α. J Biol Chem 271:17161–17166

Reich DE, Ruiz Linares A, 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

Rosqvist R, Skurnik M, Wolf-Watz H (1988) Increased virulence of Yersinia pseudotuberculosis by two independent mutations. Nature 344:522–524

Ruvolo M, Zehr S, von Dornum M, Pan D, Chang B, Lin J (1993) Mitochondrial COII sequences and modern human origins. Mol Biol Evol 10:1115–1135

Samson M, Labbe O, Mollereau C, Vassart G, Parmentier M (1996*a*) Molecular cloning and functional expression of a new human CC chemokine receptor gene. Biochemistry 35: 3362–3367

Samson M, Libert F, Doranz BJ, Rucker J, Liesnard C, Farber C-M, Saragosti S, et al (1996*b*) Resistance to HIV-1 infection in Caucasian individuals bearing mutant alleles of the CCR-5 chemokine receptor gene. Nature 382:722–725

Slatkin M, Hudson RR (1991) Pairwise comparisons of mitchondrial DNA sequences in stable and exponentially growing populations. Genetics 129:555–562

Smith MW, Dean M, Carrington M, Winkler C, Huttley G, Lomb DA, Goedert J, et al (1997) Contrasting genetic influence of CCR2 and CCR5 receptor gene variants on HIV-1 infection and disease progression. Science 277:959–965

Stoneking M, Sherry ST, Redd AJ, Vigilant L (1992) New approaches to dating suggest a recent age for the human mtDNA ancestor. Philos Trans R Soc Lond B Biol Sci 337: 167–175

Takahata N, Satta Y, Klein J (1992) Polymorphism and balancing selection at major histocompatibility complex loci. Genetics 130:925–938

Theodorou I, Meyer L, Magierowska M, Katlama C, Rouzious C, Seroco Study Group (1997) HIV-1 infection in an individual homozygous for CCR5-Δ32. Lancet 349:1219–1220

Tishkoff SA, Dietzsch E, Speed W, Pakstis AJ, Kidd JR, Cheung K, Bonné-Tamir B, et al (1996) Global patterns of linkage disequilibrium at the CD4 locus and modern human origins. Science 271:1380–1387

Vigilant L, Stoneking M, Harpending H, Hawkes K, Wilson AC (1991) African populations and the evolution of human mitochondrial DNA. Science 253:1503–1507

Vogel F, Motulsky AG (1997) Human genetics: problems and approaches, 3d ed. Springer, New York

von Haeseler A, Sajantila A, Paabo S (1996) The genetical archaeology of the human genome. Nat Genet 14:135–140

Watterson GA, Guess HA (1977) Is the most frequent allele the oldest? Theor Popul Biol 11:141–160

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

Zimmerman PA, Buckler-White A, Alkhatib G, Spalding T, Kubofcik J, Combadiere C, Weissman D, et al (1997) Inherited resistance to HIV-1 conferred by an inactivating mutation in CC chemokine receptor 5–studies in populations with contrasting clinical phenotypes, defined racial background, and quantified risk. Mol Med 3:23–26