Dating the Origin of the CCR5-Δ32 AIDS-Resistance Allele by the Coalescence of Haplotypes

J. Claiborne Stephens,1 David E. Reich,17 David B. Goldstein,17 Hyyoung 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 Giovino,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 Claustrès,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 Dean1

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

Summary

The CCR5-Δ32 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-Δ32 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-Δ32–containing ancestral haplotype to be ~700 years ago, with an estimated range of 275–1,875 years. The geographic cline of CCR5-Δ32 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.

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. 1996b). The CCR5-Δ32 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-Δ32/Δ32 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-Δ32 deletion, a high allele frequency among several Caucasian populations (Dean et al. 1996; Huang et al. 1996; Liu...
et al. 1996; Samson et al. 1996b; Martinson et al. 1997; Michael et al. 1997), and its rarity or absence in non-
Caucasian populations led to speculation that the mu-
tation occurred only once in the ancestry of the Cau-
casian ethnic group, subsequent to the continental iso-
lolation of Caucasians from African ancestors (Dean et al. 
1996; O’Brien and Dean 1997). Molecular anthropol-
ogists 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; Ruvulo et al. 1993; 
Goldstein et al. 1995; Horai et al. 1995; von Hae
seler et al. 1996). Furthermore, in attempts to explore the 
veracity of the mitochondrial “Eve Hypothesis,” con-
siderable evidence has been assembled that argues 
against the occurrence of a significant population bot-
tleneck or demographic contraction since that early di-
vergence 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 
to present frequencies of 5%–15%. Following the se-
quential selective pressures (likely on the 
CCR5-Δ32 allele) and that this haplotype was elevated by nat-
ural selective pressures (likely on the CCR5-Δ32 allele) 
to present frequencies of 5%–15%. Following the se-
lective increase, derivative modern Caucasian haplotypes 
appeared, allowing a coalescence-based estimation of the 
time required to produce the present haplotype distri-
bution. The age of that CCR5-Δ32-bearing haplotype and 
possibly the CCR5-Δ32 variant was computed, by 

<table>
<thead>
<tr>
<th>Ethnic Group</th>
<th>No. of Individuals</th>
<th>Allele Frequency</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Swedish</td>
<td>131</td>
<td>.137</td>
<td>.021</td>
</tr>
<tr>
<td>Russian</td>
<td>50</td>
<td>.136</td>
<td>.034</td>
</tr>
<tr>
<td>Estonian</td>
<td>158</td>
<td>.133</td>
<td>.019</td>
</tr>
<tr>
<td>Polish</td>
<td>30</td>
<td>.133</td>
<td>.044</td>
</tr>
<tr>
<td>Slovakian</td>
<td>30</td>
<td>.133</td>
<td>.044</td>
</tr>
<tr>
<td>Tatar</td>
<td>50</td>
<td>.120</td>
<td>.032</td>
</tr>
<tr>
<td>Australian</td>
<td>395</td>
<td>.118</td>
<td>.011</td>
</tr>
<tr>
<td>British</td>
<td>422</td>
<td>.117</td>
<td>.011</td>
</tr>
<tr>
<td>Irish</td>
<td>31</td>
<td>.113</td>
<td>.040</td>
</tr>
<tr>
<td>German</td>
<td>208</td>
<td>.108</td>
<td>.015</td>
</tr>
<tr>
<td>Czech</td>
<td>161</td>
<td>.102</td>
<td>.017</td>
</tr>
<tr>
<td>Spanish</td>
<td>56</td>
<td>.098</td>
<td>.028</td>
</tr>
<tr>
<td>Ashkenazi</td>
<td>503</td>
<td>.097</td>
<td>.009</td>
</tr>
<tr>
<td>Finn</td>
<td>195</td>
<td>.091</td>
<td>.015</td>
</tr>
<tr>
<td>French</td>
<td>230</td>
<td>.089</td>
<td>.013</td>
</tr>
<tr>
<td>Austrian</td>
<td>36</td>
<td>.089</td>
<td>.033</td>
</tr>
<tr>
<td>Danish</td>
<td>24</td>
<td>.083</td>
<td>.040</td>
</tr>
<tr>
<td>Albanian</td>
<td>73</td>
<td>.082</td>
<td>.023</td>
</tr>
<tr>
<td>Slovenian</td>
<td>110</td>
<td>.077</td>
<td>.018</td>
</tr>
<tr>
<td>Turkish</td>
<td>40</td>
<td>.063</td>
<td>.027</td>
</tr>
<tr>
<td>Italian</td>
<td>172</td>
<td>.055</td>
<td>.012</td>
</tr>
<tr>
<td>Azerbaijani</td>
<td>40</td>
<td>.050</td>
<td>.024</td>
</tr>
<tr>
<td>Bulgarian</td>
<td>29</td>
<td>.045</td>
<td>.027</td>
</tr>
<tr>
<td>Greek</td>
<td>160</td>
<td>.044</td>
<td>.011</td>
</tr>
<tr>
<td>Uzbek</td>
<td>29</td>
<td>.034</td>
<td>.024</td>
</tr>
<tr>
<td>Bulgarian Gypsy</td>
<td>47</td>
<td>.032</td>
<td>.018</td>
</tr>
<tr>
<td>Kazakh</td>
<td>50</td>
<td>.030</td>
<td>.017</td>
</tr>
<tr>
<td>Mexican</td>
<td>42</td>
<td>.024</td>
<td>.017</td>
</tr>
<tr>
<td>Uigur</td>
<td>45</td>
<td>.022</td>
<td>.016</td>
</tr>
<tr>
<td>Tuvinian</td>
<td>50</td>
<td>.020</td>
<td>.014</td>
</tr>
<tr>
<td>Georgian</td>
<td>50</td>
<td>.020</td>
<td>.000</td>
</tr>
<tr>
<td>Lebanese</td>
<td>51</td>
<td>.020</td>
<td>.000</td>
</tr>
<tr>
<td>Saudi</td>
<td>100</td>
<td>.020</td>
<td>.000</td>
</tr>
<tr>
<td>Cheyenne</td>
<td>100</td>
<td>.020</td>
<td>.000</td>
</tr>
<tr>
<td>Pima Indian</td>
<td>78</td>
<td>.020</td>
<td>.000</td>
</tr>
<tr>
<td>Pueblo Indian</td>
<td>100</td>
<td>.020</td>
<td>.000</td>
</tr>
<tr>
<td>Korean</td>
<td>50</td>
<td>.020</td>
<td>.000</td>
</tr>
<tr>
<td>Chinese</td>
<td>40</td>
<td>.020</td>
<td>.000</td>
</tr>
</tbody>
</table>

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

use of a Markov expansion, as ~700 years old (range 
275–1,875 years).

Methods

Radiation-Hybrid Mapping

Primers for the CCR1, CCR4, GAAT12D11, 
AFMB362ub9, STRL33, D3S3582, and D3S3647 
markers were used to type the GeneBridge 4 panel of 
radiation hybrids (Research Genetics) to determine cen-
trally position, and data were submitted to the radiation-
hybrid mapping service at the Whitehead Institute.
Haplotype Analysis

Individuals homozygous for CCR5-Δ32 and CEPH families carrying the CCR5-Δ32 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-Δ32 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 \( -4N_e[p \log_2 p + (1 - p) \log_2(1 - p)] \) (Kimura and Ohta 1973), which yields 6,500 generations for CCR5-Δ32, 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-Δ32 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-Δ32 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 \( -4N_e[p \log_2 p + (1 - p) \log_2(1 - p)] \) (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-Δ32 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-Δ32 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-Δ32 haplotype and then estimate the proportion of CCR5-Δ32 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

\[
P = (1 - r)^G = e^{-rG}.
\]

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

\[
G = -\ln(P)/r,
\]

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. 1996a). 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. 1A), and distances have been estimated. From a regression of STRP loci on chromosome 3 (fig. 1B), 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, \( c = 0.64 \times 0.21\% + 0.48 \times 0.72\% = 0.005 \) 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 \( \mu = 0.001 \) above, for mutation, to get \( r = 0.006 \) 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 negligible (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 \( p' = p\left(\frac{\mu w_{i1} + qw_{i2}}{w}\right) \), in which the trio \( w_{i1} \),
admixture, and virtual absence of
among northern European Caucasians, a gene-frequency
These data confirm the high frequency of
parent among Central Asian groups such as Azerbaija-
However, significant frequencies of the allele were ap-
uals, as identified in 38 ethnic groups from Europe, Asia,
(Chartl and Clark 1989). For example, if
Results
- and
- adjusting for and 5,000, respectively.

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

Table 2

<table>
<thead>
<tr>
<th>CCR5 Haplotypes Observed in Modern Caucasians</th>
<th>N (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CCR5-\Delta 32</td>
<td></td>
</tr>
<tr>
<td>CCR5-GAAT-AFMB:</td>
<td></td>
</tr>
<tr>
<td>( \Delta 32-197-215^a )</td>
<td>39 (84.8)</td>
</tr>
<tr>
<td>( \Delta 32-197-217^b )</td>
<td>3 (6.5)</td>
</tr>
<tr>
<td>( \Delta 32-193-215^a )</td>
<td>2 (4.3)</td>
</tr>
<tr>
<td>( \Delta 32-197-219^c )</td>
<td>1 (2.2)</td>
</tr>
<tr>
<td>( \Delta 32-197-213^d )</td>
<td>1 (2.2)</td>
</tr>
<tr>
<td>Total</td>
<td>46 (100)</td>
</tr>
<tr>
<td>CCR5+</td>
<td></td>
</tr>
<tr>
<td>CCR5-GAAT-AFMB:</td>
<td></td>
</tr>
<tr>
<td>+197-215</td>
<td>53 (36.3)</td>
</tr>
<tr>
<td>+197-217</td>
<td>45 (30.8)</td>
</tr>
<tr>
<td>+193-215</td>
<td>20 (13.7)</td>
</tr>
<tr>
<td>+197-219</td>
<td>2 (1.4)</td>
</tr>
<tr>
<td>+193-217</td>
<td>21 (14.4)</td>
</tr>
<tr>
<td>+191-217</td>
<td>2 (1.4)</td>
</tr>
<tr>
<td>+191-215</td>
<td>3 (2.1)</td>
</tr>
<tr>
<td>Total</td>
<td>146 (100)</td>
</tr>
</tbody>
</table>

\( ^a \) Ancestral haplotype.
\( ^b \) Recombinational origin.
\( ^c \) Either mutational or recombinational origin.
\( ^d \) Mutational origin.

\( w_{12} \), and \( \bar{w} \) is adjusted depending on whether \( CCR5-\Delta 32 \) is dominant, codominant, or recessive to wild type (Hartl and Clark 1989). For example, if \( CCR5-\Delta 32 \) is dominant, \( w_{11} = w_{12} = 1, w_{22} = 1 - s, \) and \( \bar{w} = 1 - sq^2, \) 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 \( p = 1/2N, \) for \( N = 1,000 \) 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 \( CCR5-\Delta 32 \) 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 Azerbaija-


In order to estimate the time interval that elapsed since the occurrence of the \( CCR5-\Delta 32 \) 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. 1996a). 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. 1A). Physical centimorgan distances were converted to recombination distances (in centimorgans) by use of a regression of centimorgan versus centimorgan distances computed for 13 STRP loci mapped to chromosome 3 by use of both linkage and radiation hybrids (fig. 1B).

In order to examine composite \( CCR5 \) allele–containing haplotypes, we genotyped 19 \( CCR5-\Delta 32/\Delta 32 \) homozygotes and 72 \( CCR5-++ \) homozygotes from AIDS cohorts (Dean et al. 1996) plus 20 \( CCR5-+/\Delta 32 \) heterozygotes and 17 \( CCR5-++ \) homozygotes from the CEPH mapping families, for the seven adjacent STRP loci (fig. 1A). 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 \( AFMB362w/b9 \)). These loci were mapped by use of radiation hybrids, and their recombination interval was estimated, from figure 1B, as 0.21 and 0.93 cM, respectively, from \( CCR5 \). A common \( (p = 10\%-15\%) \) missense-mutation allele (641) 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. 1A) that are at greater linkage distances (>4.1 cM) show lower or no level of linkage disequilibrium with \( CCR5 \). High linkage disequilibrium of \( CCR5-\Delta 32 \) 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 \( CCR5-\Delta 32 \)–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 recombin-
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 \( p_r \) (.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. 1B), using 13 STRP loci mapped on chromosome 3, with both linkage and radiation hybrids. The regression shows a high precision or correlation \( (r^2 = .884; \ p = 2 \times 10^{-4}; \ \text{fig. 1B}) \), although there is a modest departure in the centimorgan:centiray concordance in the actual region (175–185 cR) where the haplotype resides (see fig. 1B), 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. 1A], 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 \( \leq 100,000 \) 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. 1996b; 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 Moltusky 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 \(6.2 \times 10^{-8}\), on the assumption that it starts at 1/2\(N_e\), 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 persistence of the three-locus CCR5-Δ32 mutation 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; Cornelia 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

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
CFTR gene with postnatal female survival. Hum Genet 99: 565–572
Watterson GA, Guess HA (1977) Is the most frequent allele the oldest? Theor Popul Biol 11:141–160